

Energy Monitoring and Targeting as diagnosis;  
Applying work analysis to adapt  
a statistical change detection strategy  
using representation aiding

by

Antony Hilliard

A thesis submitted in conformity with the requirements  
for the degree of Doctor of Philosophy, Industrial Engineering

Department of Mechanical & Industrial Engineering  
University of Toronto

© Copyright by Antony Hilliard 2015



# Energy Monitoring and Targeting as diagnosis; Applying work analysis to adapt a statistical change detection strategy using representation aiding

Antony Hilliard

Doctor of Philosophy, Industrial Engineering

Department of Mechanical & Industrial Engineering  
University of Toronto

2015

## Abstract

Energy Monitoring and Targeting is a well-established business process that develops information about utility energy consumption in a business or institution. While M&T has persisted as a worthwhile energy conservation support activity, it has not been widely adopted. This dissertation explains M&T challenges in terms of diagnosing and controlling energy consumption, informed by a naturalistic field study of M&T work. A Cognitive Work Analysis of M&T identifies structures that diagnosis can search, information flows un-supported in canonical support tools, and opportunities to extend the most popular tool for M&T: Cumulative Sum of Residuals (CUSUM) charts. A design application outlines how CUSUM charts were augmented with a more contemporary statistical change detection strategy, Recursive Parameter Estimates, modified to better suit the M&T task using Representation Aiding principles. The design was experimentally evaluated in a controlled M&T synthetic task, and was shown to significantly improve diagnosis performance.



# Acknowledgments

Above all, thanks to Linda for her patience and support.

Thanks to Greg Jamieson for supporting this work in an unusual and difficult domain, and to my committee members for their guidance and thoughtful critiques.

This work could not have been completed without the contributions of field study participants, including management and staff at Energent Inc. Experimental participants would not have been available without support from energy technology professors at Humber and Seneca College.

Students at the Cognitive Engineering as well as Human Factors and Applied Statistics Labs provided design, experiment, pilot testing, data validation, and writing feedback.

This research was funded by the Association of Major Power Consumers in Ontario, the Ontario Centres of Excellence, Energent Inc., the Natural Science and Engineering Council of Canada, The Federal Economic Development Agency for Southern Ontario, and departmental support.



# Table of Contents

<b>CHAPTER 1 ENERGY MONITORING AND WORK ANALYSIS .....</b>	<b>1</b>
1.1 FOCUS: PRACTICAL ENERGY M&T CHALLENGES .....	2
1.1.1. 'Work that EMs do', and differences between jobs .....	2
1.1.2. Why focus on M&T for EMs? .....	4
1.2 SOCIETAL IMPLICATIONS OF M&T.....	5
1.2.1. Efficiency and energy gaps.....	5
1.2.2. Cognitive cost of Energy Efficiency.....	6
1.2.3. Cognitive engineering to tilt equilibrium.....	6
1.3 THEORETICAL WORK ANALYSIS & DESIGN CHALLENGES .....	7
1.3.1. Work Support Approaches.....	7
1.3.2. Invariant constraints for formative design .....	8
1.3.3. Ecological problem-solving structures.....	8
1.3.4. Tractable cost-effectiveness.....	8
1.4 STRUCTURE AND CONTRIBUTIONS OF THE DISSERTATION .....	9
1.4.1. Contribution 1: Describing the M&T task in the context of Energy Management .....	10
1.4.2. Contribution 2: Characterize M&T task in CWA framework.....	10
1.4.3. Contribution 3: Apply CWA to M&T tool design .....	11
1.4.4. Contribution 4: Controlled evaluation of M&T tool .....	12
1.5 SIGNIFICANCE OF THE CONTRIBUTIONS.....	12
<b>CHAPTER 2 DESCRIBING M&amp;T WORK AS REPORTED AND OBSERVED.....</b>	<b>13</b>
2.1 MOTIVATION: COMPARING M&T NORMS WITH A DESCRIPTIVE ACCOUNT .....	13
2.1.1. Residential field observations informing policy and design .....	13
2.1.2. Methods used to investigate M&T work .....	14
2.1.2.1. Interviews with Industrial Energy Managers.....	14
2.1.2.2. M&T Participant Observation.....	15
2.2 M&T PRACTICE, PAST AND PRESENT .....	15
2.2.1. Management Systems.....	16
2.2.2. Model-comparative monitoring .....	17
2.2.3. Good Housekeeping.....	17
2.2.4. Real-time utility meter monitoring.....	17
2.2.5. Automated control or decision support.....	18
2.2.6. First principles energy efficiency analysis .....	19
2.2.7. Discussion: Tools for which approach? .....	19
2.3 M&T WITH RECURSIVE CUMULATIVE SUM OF RESIDUALS (CUSUM).....	20
2.3.1. Origin of CUSUM method.....	20
2.3.2. Method Outline .....	21
2.3.3. Model Building .....	22
2.3.4. Model Application and Calculations .....	23
2.3.5. Interpretation.....	24
2.3.6. Problems with CUSUM Interpretation .....	24
2.3.7. CUSUM Conclusions.....	26
2.4 OBSERVATIONAL STUDY.....	26
2.4.1. Environments Observed.....	26
2.4.2. Workers.....	27
2.4.3. Methods .....	28
2.4.4. Limitations.....	29
2.5 RESULTS OF M&T OBSERVATIONS .....	29
2.5.1. Participant discoveries and learning.....	29

2.5.2.	<i>Data Records</i> .....	30
2.5.3.	<i>Energy Interpreting</i> .....	31
2.5.4.	<i>Understanding of business structure</i> .....	32
2.5.5.	<i>Model Developing</i> .....	33
2.5.6.	<i>CUSUM Interpreting</i> .....	35
2.5.7.	<i>Model Understanding</i> .....	38
2.5.8.	<i>Thought Experiments</i> .....	39
2.5.9.	<i>Social dynamics</i> .....	39
2.6	<b>DISCUSSION: PRACTICAL GAPS IN M&amp;T TOOLS</b> .....	40
2.6.1.	<i>Software needs for management vs. problem-solving</i> .....	40
2.6.2.	<i>Assessing data and models</i> .....	41
2.6.3.	<i>Comparison to practice described in literature</i> .....	42
2.6.4.	<i>Understanding and supporting diagnosis work</i> .....	43
2.6.5.	<i>Conclusion</i> .....	44

**CHAPTER 3 M&T FROM A COGNITIVE WORK ANALYSIS PERSPECTIVE.....47**

3.1	<b>MOTIVATION</b> .....	47
3.2	<b>METHODS: COGNITIVE WORK ANALYSIS</b> .....	47
3.2.1.	<i>Energy M&amp;T as Control</i> .....	48
3.2.1.1.	Goals for M&T .....	49
3.2.1.2.	Affecting the state of energy consumption.....	49
3.2.1.3.	Models of system energy performance for M&T.....	50
3.2.1.4.	Ascertaining System State in M&T .....	50
3.2.2.	<i>Method: Work Domain Analysis</i> .....	51
3.2.3.	<i>Method: Control Task Analysis</i> .....	52
3.2.4.	<i>Method: Strategy Analysis</i> .....	53
3.2.5.	<i>Analysis Methodology</i> .....	55
3.3	<b>RESULTS: WORK DOMAIN ANALYSIS</b> .....	55
3.3.1.	<i>Abstraction and existing M&amp;T knowledge bases</i> .....	55
3.3.2.	<i>Aggregation and Decomposition</i> .....	59
3.3.3.	<i>Causal/Topographical structure</i> .....	61
3.4	<b>RESULTS: CONTROL TASK ANALYSIS</b> .....	61
3.4.1.	<i>Work Situations</i> .....	61
3.4.2.	<i>Work Activities</i> .....	62
3.4.3.	<i>Decision Ladders of Energy control and Data cultivation</i> .....	65
3.5	<b>RESULTS: STRATEGIES ANALYSIS</b> .....	68
3.5.1.	<i>Overview of M&amp;T Strategies</i> .....	69
3.5.2.	<i>Strategy summaries</i> .....	70
3.5.2.1.	Energy Consumption and Cost analysis.....	71
3.5.2.2.	Reconciling Equipment Inventory with Consumption .....	71
3.5.2.3.	Consumption time-series profile pattern-detection .....	72
3.5.2.4.	Event-action times-series association.....	73
3.5.3.	<i>Condition survey Strategy</i> .....	73
3.5.4.	<i>Comparative Analysis Strategy</i> .....	76
3.6	<b>SOCIO-ORGANIZATIONAL ANALYSIS</b> .....	81
3.7	<b>DISCUSSION:</b> .....	83
3.7.1.	<i>Work Domain change and representation maintenance</i> .....	83
3.7.2.	<i>Including representation in Control Tasks</i> .....	84
3.7.3.	<i>Representation vs. productive work</i> .....	84
3.7.4.	<i>Ambiguity requires knowledge-based behavior</i> .....	85
3.7.5.	<i>Control-Theoretic perspective: sophistication versus ambiguity</i> .....	85
3.7.6.	<i>Limits of analysis</i> .....	87
3.8	<b>CONCLUSIONS FOR DESIGN</b> .....	88
3.8.1.	<i>Support for diagnosis, beyond ‘alerting’</i> .....	88
3.8.2.	<i>Making structures searchable</i> .....	88
3.8.3.	<i>Maximizing benefit of imperfect data and simple models</i> .....	90
3.8.4.	<i>What features are needed in M&amp;T software tools?</i> .....	90

## CHAPTER 4 APPLYING WORK ANALYSIS TO DEVELOP TWO NOVEL M&T DIAGNOSIS AIDS .... 91

4.1	MOTIVATION AND OPPORTUNITY .....	91
4.1.1.	<i>Collaboration with software developer</i> .....	91
4.1.2.	<i>Design Objectives</i> .....	91
4.1.3.	<i>Design Methods</i> .....	92
4.1.4.	<i>Information Requirements</i> .....	93
4.2	PROTOTYPE: MODEL SUMMARY SHEET .....	93
4.3	PROTOTYPE: RECURSIVE PARAMETER ESTIMATES CHARTS.....	96
4.3.1.	<i>Statistical time-series change detection</i> .....	96
4.3.2.	<i>Origin of Recursive Parameter Estimates Method</i> .....	96
4.3.3.	<i>Method Outline</i> .....	97
4.3.4.	<i>Statistical weaknesses in Recursive Estimates Charts</i> .....	98
4.3.5.	<i>Perceptual ambiguities in Recursive Estimates Charts</i> .....	99
4.3.6.	<i>Utility of test statistics in Recursive Estimates Charts</i> .....	100
4.3.7.	<i>Disambiguating Recursive Estimates Charts</i> .....	101
4.3.7.1.	Exponentially Weighted Memory .....	101
4.3.7.2.	Meta-Information .....	103
4.3.7.3.	Relative Scaling .....	104
4.3.8.	<i>Interpreting Modified RE Charts</i> .....	104
4.3.9.	<i>Annotating Modified RE Charts</i> .....	106
4.4	EXAMPLE RECURSIVE ESTIMATES CHARTS.....	106
4.4.1.	<i>Synthetic data</i> .....	106
4.4.2.	<i>Hospital example</i> .....	108
4.5	CONCLUSION: IMPLEMENTATION AND EVALUATION .....	111
4.5.1.	<i>Remaining weaknesses of Recursive Estimates Charts</i> .....	111
4.5.2.	<i>Extra value from same model</i> .....	112
4.5.3.	<i>Time-series Representation Aiding</i> .....	113
4.5.4.	<i>Heuristic and Usability evaluation</i> .....	114

## CHAPTER 5 CONTROLLED EVALUATION OF TIME-SERIES CHANGE DIAGNOSIS SUPPORT .... 115

5.1	MOTIVATION: M&T NEVER STUDIED AS CONTROLLED ENVIRONMENT .....	115
5.1.1.	<i>Goal 1: Study M&amp;T performance with standard tools</i> .....	115
5.1.2.	<i>Goal 2: Assess new normative diagnosis strategy support</i> .....	116
5.1.3.	<i>Experimental Hypotheses</i> .....	116
5.2	METHOD: SYNTHETIC M&T TASK AND EXPERIMENTAL DESIGN .....	116
5.2.1.	<i>Stimulus development</i> .....	116
5.2.2.	<i>Scenario development</i> .....	119
5.2.3.	<i>Experimental design</i> .....	121
5.2.4.	<i>Response forms and equipment</i> .....	122
5.2.5.	<i>Participants</i> .....	122
5.2.6.	<i>Administration method</i> .....	123
5.2.7.	<i>Data Interpretation and Entry</i> .....	123
5.2.8.	<i>Data Scoring</i> .....	126
5.2.9.	<i>Data Categorization and Measures</i> .....	127
5.2.10.	<i>Questionnaire</i> .....	131
5.2.11.	<i>Statistical Methods</i> .....	131
5.3	RESULTS, PILOT EXPERIMENT I.....	131
5.3.1.	<i>Participation</i> .....	132
5.3.2.	<i>Response time and types</i> .....	132
5.3.3.	<i>By-Participant performance and Interface effects</i> .....	133
5.3.3.1.	Detection Interface Effects.....	135
5.3.3.2.	Diagnosis Interface Effects .....	135
5.3.4.	<i>Stimulus Effects</i> .....	135
5.3.4.1.	By-Change Properties Detection effects .....	135
5.3.4.2.	By-Change Properties Diagnosis effects .....	136
5.4	PILOT EXPERIMENT I DISCUSSION.....	136

5.5	RESULTS, EXPERIMENT II.....	137
5.5.1.	<i>Participation</i> .....	137
5.5.2.	<i>Response types and time</i> .....	137
5.5.2.1.	Response Windows.....	138
5.5.2.2.	Response locations.....	139
5.5.2.3.	Time taken.....	140
5.5.3.	<i>By-Participant performance and Interface effects</i> .....	141
5.5.3.1.	Detection Interface Effects.....	142
5.5.3.2.	Diagnosis Interface Effects.....	144
5.5.3.3.	Overall Interface effects.....	144
5.5.4.	<i>Stimulus Effects</i> .....	145
5.5.4.1.	By-Change Properties Detection effects.....	146
5.5.4.2.	By-Change Properties Diagnosis effects.....	148
5.5.5.	<i>Other influences on Experiment II results</i> .....	149
5.5.5.1.	Data Entry validation.....	149
5.5.5.2.	Scoring rules.....	149
5.5.5.3.	Learning effects.....	150
5.5.5.4.	Marked visual aids.....	151
5.5.5.5.	Confidence in responses.....	151
5.5.5.6.	Propensity to attempt diagnosis with each change type.....	151
5.5.6.	<i>Participant Feedback</i> .....	157
5.5.6.1.	Questionnaire.....	157
5.5.6.2.	Comments.....	158
5.6	EXPERIMENT DISCUSSION.....	159
5.6.1.	<i>Evaluating performance with standard (CUSUM) tools</i> .....	159
5.6.2.	<i>Evaluating support of Recursive Estimates-based strategies</i> .....	160
5.6.3.	<i>Scenario effects on detection and diagnosis</i> .....	161
5.6.4.	<i>Change type effects on detection and diagnosis</i> .....	162
5.6.5.	<i>Support for experimental hypotheses</i> .....	164
5.6.6.	<i>Is the experimental validity sufficient?</i> .....	165
5.7	CONCLUSIONS FOR M&T DETECTION AND DIAGNOSIS.....	167

**CHAPTER 6 DISCUSSION AND FUTURE WORK.....171**

6.1	CONTRIBUTIONS.....	173
6.1.1.	<i>Naturalistic description of M&amp;T behavior</i> .....	173
6.1.2.	<i>Cognitive Work Analysis of M&amp;T</i> .....	174
6.1.3.	<i>Model summary sheet &amp; modified Recursive Estimates</i> .....	175
6.1.4.	<i>Controlled M&amp;T task experiment</i> .....	176
6.2	INCIDENTAL CONTRIBUTIONS.....	177
6.3	LIMITATIONS AND FUTURE WORK.....	177
6.3.1.	<i>Field studies of true expert M&amp;T behavior</i> .....	177
6.3.2.	<i>Future experimental work</i> .....	178
6.3.3.	<i>CWA Future work</i> .....	179
6.3.4.	<i>Extending RE charts for energy performance diagnosis</i> .....	180
6.3.5.	<i>Multi-strategy “ecological” M&amp;T information systems</i> .....	180
6.4	CONCLUSIONS.....	181

# Table of Figures

Figure 1 - Conceptual Venn diagram of Monitoring & Targeting (M&T) and its overlap with other Energy Management and business operation activities.....	3
Figure 2 - M&T as identifying energy waste in a business .....	4
Figure 3 - Outline of dissertation chapters and how findings from one section were applied in others. At top, chapters and associated contributions. At bottom, three aspects from each chapter.....	9
Figure 4 - Data transformation steps in CUSUM algorithm. Consumption (y) and driver variables (x) at top, processed time-series charts at right. ....	21
Figure 5 - CUSUM chart illustrating how a single business change (at top center on 1 Feb, 11) can produce an wavy, ambiguous CUSUM chart. Overlaid straight lines show how standard CUSUM interpretation rules might be (mis)applied. Such changes can be associated with an intermittent operation mode (e.g. excess fresh-air ventilation), or a true permanent change (e.g. a hole in a wall or window). ....	25
Figure 6 - Rough proportion of M&T task time analysts and energy specialists spent interpreting energy data, assessing data or models, and navigating the M&T information system. Coding not validated. ....	30
Figure 7 - Example of purposeful model design for a heating system (as a block diagram, top). Models (at left) could include (checkmark) measures of certain processes to isolate furnace combustion and/or building envelope insulation. A simpler model (bottom) might omit measures of indoor temperature so that the resulting CUSUM chart would respond to turning down the thermostat. ....	34
Figure 8 - Example of a CUSUM chart illustrating how a large system change (top left) creates a diagonal feature (center) that obscures smaller changes that would be more perceptible otherwise .....	36
Figure 9 - Example of a CUSUM chart showing multiple overlapping changes. From this chart it is not clear whether changes in CUSUM slope (December, April, September, June, December) show unrelated changes or fewer common persistent changes. The same data is processed with RE charts in Section 4.4.1. ....	37
Figure 10 - Illustration of different explanations of energy models offered by Energy analysts (top), the Site A factory energy specialist (middle), and Site B hospital energy specialist (bottom). ....	38
Figure 11 - Most published Cognitive Work Analyses have not described Strategies (Hassall & Sanderson, 2014, p. 221).....	48
Figure 12 - Examples of Decision Ladder annotation used in Control Task Analysis. At left, normative information-processing steps (lines) between knowledge states (circles). At right, (unrelated) examples of notation for associative leaps and information-processing shunts. Adapted from (Lintern, 2009, p. 65).....	53
Figure 13 - Some typical functional activities of Energy Management (Hilliard et al., 2009). Energy M&T represented mostly by "Consumption monitoring & modeling" with some "Target setting". EMO stands for Energy Management Opportunity. Arrows represent information flow between activities (Table 6) .....	63
Figure 14 - Linked decision ladders for two M&T work functions: Cultivating Data & Models (left), and Controlling Energy Costs (right). Ovals represent states of knowledge, described in Appendix A.1. Arrows represent information processing steps. ....	67
Figure 15 - Information Flow Map for Comparative Analysis strategy, transforming observations (bottom) to a system state indicator (top). Energy models are maintained and updated from time to time (center). Darker shading indicates elements that were poorly documented and less observable to field study participants .....	79

Figure 16 - Comparative Analysis strategy as a feedback control loop for a System subject to Disturbances. (Adapted from Jamieson & Vicente, 2005) .....86

Figure 17 - Examples of how linear regression models used for CUSUM comparative analysis were presented to end-users. This is representative of three M&T software systems I observed in the course of participant observation and field study.....94

Figure 18 - Model Summary Sheet Prototype, as delivered to client for implementation. First two pages are fixed-format, third page can expand as necessary to accommodate models with more driver variables and associated parameters.....95

Figure 19 – Data flow diagram for calculating RE charts. Existing CUSUM chart data flow (Figure 4) shown in light grey. ....98

Figure 20 - Synthetic data for a three-term model: Intercept, Heating Degree Days, and Precipitation. Three changes introduced, first to Intercept, second to HDD, and last to Precipitation. RE plots scaled to indicate statistical significance vs. test statistic (red). Contrast with modified RE chart below in Figure 21. ....102

Figure 21 – Same data as Figure 20 in modified RE plots with exponential decay at 60-month and 2-month time constants. Plots scaled analogously to change time and size. The steady height of the middle 2-month (grey) HDD chart shows that the change in heating process efficiency that occurred December 10<sup>th</sup> persisted over two summers with a steady, comparable effect.....102

Figure 22 - Stereotypical behavior of RE charts at long (L) and short (S) exponentially decayed memory, compared to CUSUM charts and the actual underlying change reflected in model intercept (baseload), or a parameter (driver sensitivity). Interpretation rules 1 and 2 apply consistently to baseload and driver changes.....105

Figure 23 - Example of CUSUM chart from Figure 9, with RE charts developed from same model, and a superimposed “solution”. The linear model has an intercept (baseload) plus two parameters, seasonal heating plus intermittent precipitation. RE charts are plotted at two exponential decay timescales: 2-year (black line) and 2-month (grey shaded). ....107

Figure 24 - CUSUM Chart, Healthcare Institution Natural Gas Consumption, Fall 2010-2011. From top to bottom: Consumption (thin black) and Modeled consumption (thick blue) charts. Control Chart showing over/underconsumption. Finally, CUSUM chart showing times of three suspected changes (A,B,C).....109

Figure 25 - RE Chart, Hospital Natural Gas Consumption, Fall 2010-2011. RE charts for the same model and period as Figure 1. Each chart indicates change in a model parameter: baseload (Intercept), Weekday, and Heating Degree Day. Each chart shows two exponential memory decay factors: Grey bars indicate a fast-responding, 2-month time constant. Black lines indicate a slow-responding 60-month time constant.....110

Figure 26 - Diagnostic search starting points with CUSUM charts (left) are in terms of over/under consumption in time at a particular utility meter. RE charts expand diagnostic search in terms of energy model drivers often representative of more abstract system structure (right) .....113

Figure 27 - One of the three years of data used for the Experiment II stimulus, including synthetic "Generator Hours" variable, Heating Degree Day (HDD) conversion of local temperature, and workday binary measure. ....117

Figure 28 – Accumulated Evidence of four persistent changes, 4A (-10% workday), 4B (+10% baseload), 4C (-10% heating), and 4D (+20%, workday) superimpose (bottom) and sum to create the Experiment II Scenario 4 CUSUM chart (top). “Generator” performance did not change in this scenario.....121

Figure 29 - Example response from Experiment II, Participant 209, Order C, Trial 10, Scenario 5, G<sub>1..5</sub>. ....125

Figure 30 – Taxonomy for M&T experiment data, described as a data analysis decision flow chart. Participant responses (R) are scored against true changes (TC) in scenarios according to nested detection (H, FA) and diagnosis criteria (AD, ND)..... 128

Figure 31 - Time taken in Experiment I to complete each booklet of 5 scenarios ( $N=32$ ), according to first booklet (outer sort) and whether the scenarios were displayed with CUSUM-only or CUSUM+RE charts (inner sort)..... 133

Figure 32 – Experiment I Detection and Diagnosis summary, according to “Likely” scoring rule, aggregated by CUSUM-only or CUSUM+RE experimental condition. Top row are counts of R, AD, H, RD. Bottom row are rate-normalized according to Table 17. .... 134

Figure 33 - Width of response windows, for all participants/scenarios/conditions. A response width of 100 days corresponds to a 1.9cm line drawn on a scenario chart. .... 138

Figure 34 – Pearson correlations ( $N=1329$ ) between response window width (BoxWidth), whether the response overlapped a change (InBox), the indicated confidence in the response, and the Interface condition. .... 139

Figure 35 - Location of response marks on response sheet, by interface condition (CUSUM or CUSUM + RE) .... 140

Figure 36 - Time taken in Experiment II to complete each booklet of 5 scenarios ( $N=66$ ), according to first booklet (outer sort) and whether the scenarios were displayed with CUSUM-only or CUSUM+RE charts (inner sort)..... 141

Figure 37 - Detection and Diagnosis summary, according to “Likely” scoring rule, aggregated by CUSUM-only or CUSUM+RE experimental condition. Top row are counts of R, AD, H, RD. Bottom row are rate-normalized according to Table 17. \*\* =  $p < .005$ , \*\*\* =  $p < .001$  ..... 143

Figure 38 - Detection and diagnosis rates of each of 13 changes (Labeled by scenario, 1A to 5A). Scored using "likely" rule, combining changes from Normal  $G_{1..5}$  and Inverted  $G_{5..10}$  scenario sets. At top, proportion of participants who detected each change (pH), at bottom right diagnosis rate (RDr). Color indicates change cause (Table 13, Blue = Baseload, Orange = HDD, Grey = Weekday, Black = Generator). .... 147

Figure 39 - Proportions of change Detections (pH) and associated Right Diagnosis rates (RDr), aggregated by change Cause/Type and Interface condition. Change Causes are A (Baseload), B (Workday), C (Heating), and D (Generator). .... 148

Figure 40 - Histogram comparing response- change pairs scored by "InBox" and "Likely" rules, according to discrepancy between marked "X" and the true date of the matched change (in scenario days, 5 days = ~1mm). Darker shaded sections at left were scored by both rules. Lighter shaded scores at right are late responses only matched by "Likely" rule..... 150

Figure 41 – For all hits with an attempted diagnosis (HD), confusions between attempted diagnosis and true change causes, as a percentage of total HD. Outside loops indicate right diagnoses (RD). Perfect performance would be 25% of HD being right diagnoses of each change type. Based on data from Table 22. .... 153

Figure 42 - For Hits, how often change types were rightly diagnosed (RDr) correlates with the by-cause proportion of diagnosis attempts (HDs). Data aggregated by type of change Hit by response (A,B,C,D). Scale lines indicate chance (1 in 4) performance. .... 155

Figure 43 - For each change type Hit, the percentage participants ( $n=33$ ) rightly diagnose (RDr) in a scenario booklet ( $n=264$ ) correlates with the proportion of diagnosis attempts (HDs). Scale lines indicate chance (1 in 4) performance. Data points are jittered to show density. Local least-squares trend lines fit by LOESS. .... 156

Figure 44 – Fully- or Under-constrained metering system (where utility meter consistency cannot be checked) .... 200

Figure 45 - Over constrained metering system (can cross-check meter consistency aka. 'Unaccounted electricity') 200

Figure 46 - Some potential inferences (dashed lines) relevant to interpreting system state from energy data. Inconsistencies in information processing are what the work function of "Cultivating Data & Models" seeks to understand and minimize. ....204

Figure 47 – Elements of Decision Ladder (Figure 14) active in "Automatic" Monitoring & Targeting subtask. Three states of knowledge are relevant: Alerted, Observations, and Desired State. ....205

Figure 48 – Elements of Decision Ladder (Figure 14) active in “Condition survey” subtask. Relevant states of knowledge: Alerted, Observations, Task, Procedure. ....206

Figure 49 – Elements of Decision Ladder (Figure 14) active in “Energy Audit” subtask .....207

Figure 50 – Elements of Decision Ladder (Figure 14) active in an “Energy Performance Analytics” subtask .....208

Figure 51 – Residual inspection plots for model MELogit.QIHR.19, explaining likelihood of a response being a hit. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.....236

Figure 52 – Residual inspection plots for mixed-effects attempted diagnosis rate model MEADr.glmer17, Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown. ....240

Figure 53 – Residual inspection plots for model MELogit.QIRDr16, explaining likelihood of a hit with diagnosis being right. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.....246

Figure 54 – Residual inspection plots for model MELogit.QICR16, explaining likelihood of a response being completely correct. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.....251

Figure 55 – Residual inspection plots for model MELogit.TfRDr.19, explaining likelihood of a hit diagnosis being right. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown. ....256

Figure 56 – Residual Inspection plots for model MELogit.TfCRr.17, explaining likelihood of a response being completely correct. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.....262

Figure 57 - Residual inspection plot for Model MELogitQIDet9, explaining likelihood of a change being detected. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.....269

Figure 58 - Residual inspection plot for Model MELogitQIDet13, explaining likelihood of a change being detected. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.....269

Figure 59- Residual inspection plot for Model MELogitQIDiag10, explaining likelihood of a participant rightly diagnosing a change that they detected and attempted a diagnosis for. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown. ....276

Figure 60 - Pearson Correlations and scatter plots comparing "InBox" and "Likely" scoring rules for Detection (left) and Diagnosis (right). Scores are aggregated by-participant. ....278

Figure 61 - By-Response (n=1317) correlation between indicated confidence in the change being new, and whether response Hit a true change. ....280

Figure 62 - By-Response (n=463) correlation between indicated confidence in the change being new, and whether Hits with Diagnosis attempts were Rightly Diagnosed.....280

Figure 63 - Questionnaire responses for all participants ( $N = 33$ ), on questions whether each Interface was Easy, showed When, showed What, was Informative, & was Confusing. Pearson correlations shown in top right..... 288

Figure 64 – Participant total ( $n=33$ ) counts of Responses, False Alarms, Hits, Hits w/ Diagnosis, and Right Diagnoses, by Interface condition. Scenarios combine 5 normal  $G_{1..5}$  and inverted  $G_{6..10}$  TrueScenarios. Number of changes plotted as + symbols: two in scenario 1, two in scenarios 2,3, three in 4, and one in 5..... 291

Figure 65 – Average ( $n=33$ ) ratios of Detection and Diagnosis measures, by Interface type and five Scenarios. From top: Hit, False Alarm, Attempted Diagnosis, Hit w/ Diagnosis, Right Diagnosis, and Correct Response ratios described in Table 17..... 292

# Table of Tables

Table 1 - Participating institutions and workers in observational field study .....	28
Table 2 - Comparing context that might be considered by colleagues from different business areas. Such varying perspectives will be categorized in Work Domain Analysis (Section 3.3) .....	41
Table 3 - Three perspectives on how energy can be controlled in organizations, with respect to Work Domain (Section 3.2.2) terms and social activities.....	50
Table 4 - An Abstraction - Decomposition space for Energy Management in large enterprises. Functional purposes are those of the business, not 'save energy'. Abstract Functions, Values/Priorities limited to energy-related, and only energy-relevant equipment included.....	58
Table 5 - How utility energy consumption can be dis-aggregated at different levels of abstraction .....	60
Table 6 - Example information flows between energy management activities (Hilliard et al., 2009). Work activities of Figure 13 are shown as topmost diagonal cells and information flows (as questions) in the tabular intersections between functions. ....	64
Table 7 - Six M&T situation assessment strategies, categorized by whether they need external data and model maintenance (left) or can be recognition-based (right).....	69
Table 8 - Six M&T situation assessment strategies, compared in terms of properties that may be relevant to strategy switching trade-offs .....	70
Table 9 –States of knowledge specific to Condition Survey strategy. This strategy can be conducted with mental models alone, if the actor can recognize stereotypical signs of energy waste (e.g. hot air drafts) or site-specific condition or operation problems .....	75
Table 10 - Specific states of knowledge for Comparative Analysis strategy. Unlike the Condition survey strategy in Table 9, Comparative Analysis requires cultivated data and models to inform energy cost control. ....	78
Table 11 - A conceptual mapping of how well three social organizations might support the six strategies presented above. Filled circles represent the author’s judgment of most suitable, empty least suitable.....	82
Table 12 – Experiment II Design of Stimulus Drivers and Parameters. ....	118
Table 13 – Properties (at left) of all changes in Experiment II Scenarios (at right). Leading changes appear first (labeled with alphanumeric _A) in the scenario. Large changes are double small changes. Type/Cause refers to which parameter changed (A..D). Accumulated evidence is a product of change size, driver variables, and intervening time between changes. Change 2A/7A was not scored as it served only to obscure subsequent changes (2B/7B and 2C/7C).....	120
Table 14 - Experimental Design of Blocks and associated Treatment and Scenario group Order.....	122
Table 15 - Response $\diamond$ Change scoring rules, recursively applied for each trial until no “best” pairings left.....	126
Table 16 - Taxonomy for M&T experiment data, described as a cross-tabulation of responses matched to true changes. Detection (H, FA) and Diagnosis (AD, ND) categories overlap. Misses are not shown in this table, and Correct Rejections do not apply to this experiment. See Table 22 for experimental data in this format. ....	129
Table 17 - Data Analysis ratios of categories shown in Table 16 and Figure 30. These ratios can 1) correct for and/or 2) quantify participants’ propensity to a) respond (detection) and/or b) identify a cause (diagnosis). Ratios can also describe the proportions of attempted diagnoses or hits broken down by change type/cause. ....	129

Table 18 - Data Aggregation Levels used in analysis of Experiment II data. Data becomes more aggregated from left to right (by-participant) and from top to bottom (by-experimental-stimulus). .....	130
Table 19 - Experiment 1 Participants by experimental block Order and School. Asterisk* indicates one participant (106) removed due to misuse of RE charts.....	132
Table 20 - Experiment 2 Participants by experimental block Order and School.....	137
Table 21 - Response locations for all participants ( $n = 33$ ), all scenarios ( $s = 5$ ) in each experimental condition (CUSUM only, or CUSUM+RE). Charts are numbered as they appear on the response forms, from top to bottom. ....	140
Table 22 - Response counts, cross-tabulated by Marked Cause (including non-diagnosis-attempts N/A), by Interface condition, and by True Cause of Hit according to “Likely” scoring rule (plus False Alarms N/A). Extra-large 13th change 2A / 7A omitted. Bolded diagonals represent Correct Responses. Totals not shown.....	152
Table 23 - Proportion of Hits matched against True Causes of each change type. Same data as from Table 22. Change 2A/7A omitted. Total by-diagnosis proportions (HDs) of Marked Causes for Hits with Diagnosis summarized at right. ....	154
Table 24 - Proportion of Rightly Diagnosed hits with diagnosis (RDr) compared with tendency to attempt each change diagnosis type (HDs) accompanying a hit. Summarizes Figure 39 and plotted in Figure 42.....	155
Table 25 Questionnaire Summary Statistics, for all participants ( $n=33$ ). Score of 3 is "Un-decided". .....	158
Table 26 - Dimensions on which experimental work described in this dissertation could be extended .....	178
Table 27 - Annotation of states of knowledge for "Control Energy Costs" work function shown in Figure 14.....	201
Table 28 - Annotation of states of knowledge for "Cultivate Data & Models" work function shown in Figure 14. .	202
Table 29 – Experiment I Response locations for all participants ( $n=18$ ), all scenarios ( $s=5$ ) in each experimental condition (CUSUM only, or CUSUM+RE). Charts are numbered as they appeared on the response forms, from top to bottom.....	219
Table 30 - Inter-Rater Reliability test of Experimental II data coding. 127 samples coded by 2 raters. ....	277
Table 31 - Detection Hits and False Alarms (“Likely” scoring rule), by Scenario (aggregating $G_{1..5}$ and $G_{5..10}$ scenario sets). Mean and Standard Deviation summaries shown. ....	289
Table 32 – Diagnosis Marked Causes and Right Diagnoses (“Likely” scoring rule), by Scenario (aggregating $G_a$ and $G_b$ scenario sets, $n=33$ ). Average of performance in C and C+R interface condition.....	293
Table 33 - Scenario definitions for Experiment II, by number and description (left). Changes (at right) are identified by scenario and number (e.g. 1A, 1B), and defined by their onset time (day of dataset), magnitude of parameter change, uninterrupted duration of influence, and accumulated evidence. ....	295

# List of Appendices

<b>APPENDIX A</b>	<b>COGNITIVE WORK ANALYSIS OF ENERGY M&amp;T ADDENDA .....</b>	<b>197</b>
A.1	COMPARING M&T TO OTHER HUMAN FACTORS DOMAINS .....	197
A.1.1	<i>Nuclear Power</i> .....	197
A.1.2	<i>Aviation</i> .....	197
A.1.3	<i>Safety / Risk management</i> .....	198
A.1.4	<i>Quality Control</i> .....	198
A.2	SOME SENSITIZING CONCEPTS FOR CWA OF M&T .....	198
A.3	CAUSAL / TOPOGRAPHIC STRUCTURE IN M&T .....	199
A.4	DECISION LADDER ANNOTATIONS .....	200
A.5	M&T DATA INFERENCE PROBLEMS .....	203
A.6	INSTANTIATED DECISION LADDERS.....	205
A.6.1	<i>“Automatic” Monitoring &amp; Targeting</i> .....	205
A.6.2	<i>Condition survey</i> .....	206
A.6.3	<i>Energy Efficiency Audit</i> .....	207
A.6.4	<i>Energy Performance Analytics</i> .....	208
A.7	SWITCHES BETWEEN M&T STRATEGIES .....	208
<b>APPENDIX B</b>	<b>STATISTICAL CHANGE DETECTION DISCUSSION .....</b>	<b>211</b>
B.1	TEST STATISTICS IN CUSUM CHARTS.....	211
B.1.1	<i>Change detection</i> .....	211
B.2	‘ECOLOGICAL’ INFORMATION SYSTEM FEATURES FOR M&T SUPPORT.....	211
B.2.1	<i>Support units compatible with physical sampling</i> .....	211
B.2.2	<i>Support contextual time reference frames not just calendar dates</i> .....	213
B.2.3	<i>Supports social engagement?</i> .....	213
B.2.4	<i>Forecasting and thought experiments?</i> .....	213
<b>APPENDIX C</b>	<b>EXPERIMENT I STATISTICAL OUTPUT .....</b>	<b>215</b>
C.1	EXPERIMENTAL DESIGN .....	215
C.1.1	<i>Model Training Data</i> .....	215
C.2	DATA SUMMARIES .....	215
C.2.1	<i>Performance Data, Aggregated by-Participant</i> .....	215
C.2.2	<i>Performance Data, by-Participant and Interface</i> .....	216
C.3	RESPONSES.....	217
C.3.1	<i>Task Time</i> .....	217
C.3.2	<i>Chart Location</i> .....	219
C.4	BY-PARTICIPANT PERFORMANCE.....	219
C.4.1	<i>Detection</i> .....	219
C.4.2	<i>Diagnosis</i> .....	221
C.5	BY-CHANGE TYPE PERFORMANCE .....	221
C.5.1	<i>Detection</i> .....	221
C.5.2	<i>Diagnosis</i> .....	224
<b>APPENDIX D</b>	<b>EXPERIMENT II STATISTICAL OUTPUT .....</b>	<b>229</b>
D.1	EXPERIMENTAL DESIGN .....	229
D.1.1	<i>Model training data</i> .....	229
D.2	DATA SUMMARIES FROM EXPERIMENT II.....	229
D.2.1	<i>Performance Data, Aggregated by-Participant</i> .....	229
D.2.2	<i>Performance Data, by Participant and Interface</i> .....	230
D.2.3	<i>Questionnaire Data, by Interface</i> .....	231
D.3	RESPONSE MODES.....	232
D.3.1	<i>Task Time</i> .....	232

D.4	BY-PARTICIPANT M&T PERFORMANCE .....	233
D.4.1	<i>Interface and Order Effects on Detection</i> .....	233
D.4.2	<i>Interface Effect on Attempted Diagnoses Mixed Effect Logit Models</i> .....	236
D.4.3	<i>Interface Effect on Diagnosis Accuracy Mixed Effect Logit Model</i> .....	240
D.4.4	<i>By-Interface Overall Performance Effect Mixed Effect Logit Models</i> .....	246
D.5	BY-SCENARIO PERFORMANCE.....	251
D.5.1	<i>Scenario Effects on Detection</i> .....	251
D.5.2	<i>Scenario Effect on Diagnosis Mixed Effect Logit Models</i> .....	253
D.5.3	<i>Contrasting Scenario Diagnosis Performance</i> .....	257
D.5.4	<i>Scenario Performance Difference Mixed Effect Logit Models</i> .....	258
D.6	BY-CHANGE STIMULUS EFFECTS .....	263
D.6.1	<i>Stimulus Detection Effects</i> .....	263
D.6.2	<i>Stimulus Diagnosis Effects</i> .....	270
D.7	OUTSIDE INFLUENCES / ASSUMPTION CHECKING .....	276
D.7.1	<i>Data Entry Validation</i> .....	276
D.7.2	<i>Comparing Scoring Rules –Main effects</i> .....	277
D.7.3	<i>Confidence Correlations</i> .....	279
D.7.4	<i>Comparing rates of Attempted Diagnosis</i> .....	280
D.7.4.1	Proportion of Hits and Diagnosis Attempts by Interface condition.....	281
D.7.4.2	Cross-Tabulating Misdiagnoses .....	283
D.7.4.3	Proportion of Diagnosis Attempts between Hits and False Alarms.....	286
D.7.4.4	Correlation between Diagnosis performance and Attempt tendency, by Interface.....	287
D.7.5	<i>Participant Feedback</i> .....	287
D.7.5.1	Questionnaire Correlations.....	287
<b>APPENDIX E SUPPLEMENTARY ANALYSES .....</b>		<b>289</b>
E.1	EXPERIMENT II BY-SCENARIO RESULTS .....	289
E.1.1	<i>Detection Scenario Effects</i> .....	289
E.1.2	<i>Diagnosis Scenario Effects</i> .....	293
E.1.3	<i>Overall Scenario effects</i> .....	294
<b>APPENDIX F EXPERIMENTAL MATERIALS.....</b>		<b>295</b>
F.1	EXPERIMENT II SCENARIO DEFINITIONS .....	295
F.2	EXPERIMENT II INSTRUCTIONS.....	296
F.3	SAMPLE EXPERIMENT II RESPONSE FORM .....	298
F.4	SAMPLE EXPERIMENT II QUESTIONNAIRE .....	299
F.5	EXPERIMENTAL BOOKLETS .....	300



# Glossary

This dissertation largely adopts terminology from the Cognitive Work Analysis theoretic tradition, described elsewhere (Vicente, 1999, p. 3). Other terms used in this dissertation are:

**Assessing:** Determining the internal consistency and informativeness of a cue.

**Business system structure:** Invariants that determine business system behavior. Can include physical equipment, automatic controls, or routine management processes.

**CUSUM - Cumulative Sum of Residuals:** A statistical method, analogous to quality control charts (Page, 1961).

**EM - Energy Management:** Managerial work of developing energy accounting, forecasting, and control processes within a business organization.

**EMIS – Energy Management Information System:** An information system to support M&T, M&V, or other Energy Management work

**EMRS – Energy Management Rule System:** An expert system rule-based decision aid to optimize automated or routine operation decisions. One approach to EMIS.

**Interpreting:** Determining the implications of a cue for a task-relevant judgement.

**M&T - Monitoring and Targeting:** A task of assessing and controlling the utility energy consumption performance of a system. Part of energy management work.

**M&V - Measurement and Verification:** A task to quantify known changes in utility energy consumption of a system, often for contractual use. Part of energy management work.



# Chapter 1

## Energy Monitoring and Work Analysis

This dissertation is a study, analysis, and development project motivated by a poorly understood and often overlooked task: monitoring and targeting (M&T) energy end-use. M&T is part of managing efficient energy consumption in businesses. Effectively managing energy can save 5-10% on utility bills by operation & maintenance changes alone (Carbon Trust, 2007; Therkelsen, McKane, Sabouni, Evans, & Scheihing, 2013), which for heavy industry can amount to substantial cost savings.

M&T is interesting to study for practical, academic, and social reasons. Practical challenges have persisted since M&T was introduced in the 1980s; distinguishing energy waste from justified consumption in dynamic, disturbance-prone businesses is difficult. M&T has interesting characteristics for Human Factors Engineering methods, as well; it is a diagnosis task of an information-dense cue (utility consumption) whose interpretation requires a wide range of referents, from the physical appearance of equipment to sophisticated statistical analyses. Furthermore M&T has important social implications. Climate change, the most pressing challenge facing civilization, requires rapid and drastic reduction in carbon emissions. Energy efficiency will be a key part of de-carbonization. Operating and maintaining virtually every energy-consuming process more efficiently will entail significant cognitive work and require many more workers to act mindfully of energy efficiency.

The four research questions interrogated in this dissertation are: a) what barriers to effective work occur in M&T practice? b) how can M&T challenges be described in Cognitive Engineering constructs? c) what information system features should address practical M&T barriers and improve M&T task performance? and d) can M&T barriers be replicated and assessed in a synthetic M&T task? This chapter introduces the M&T domain in general terms, with an emphasis first on practical challenges that this dissertation addresses, and secondly on theoretic work analysis issues that M&T exposes. The introduction concludes with the four main contributions of the dissertation and an outline of the chapter structure.

## 1.1 Focus: Practical Energy M&T challenges

M&T is a challenging task, and is an essential part of energy management work. Energy management applies engineering, financial, and organizational principles to help businesses make sound decisions about energy purchase, use, and control (Hooke, Landry, & Hart, 2004). This involves "linking energy to un-transparent costs, with limited time and resources" (Carbon Trust, 2007, p. 15). Following the 1974 energy crisis, government supported energy management practice across eight industrial sectors, where participating businesses reduced energy costs by 4-18% (Gotel & Hale, 1989, p. 29). But widespread M&T adoption did not occur, and practitioners found "energy management still fails to receive the attention and commitment from senior management in industry that it deserves... [why] energy efficiency should receive such scant attention is a matter of conjecture" (Harris, 1989, p. 7). In the intervening decades, despite continued promotion and subsidies<sup>1</sup>, poor energy management and uneconomic energy consumption remains widespread (Shiple & Elliott, 2006). Even very energy-intensive industries in the most developed countries can improve; a study of steelworks in Sweden found only 40% of mills and 25% of foundries successfully managed energy (Thollander & Ottosson, 2010). As the authors note, it is concerning that even in the 21<sup>st</sup> century some of the clearest opportunities for energy efficiency are not being effectively pursued.

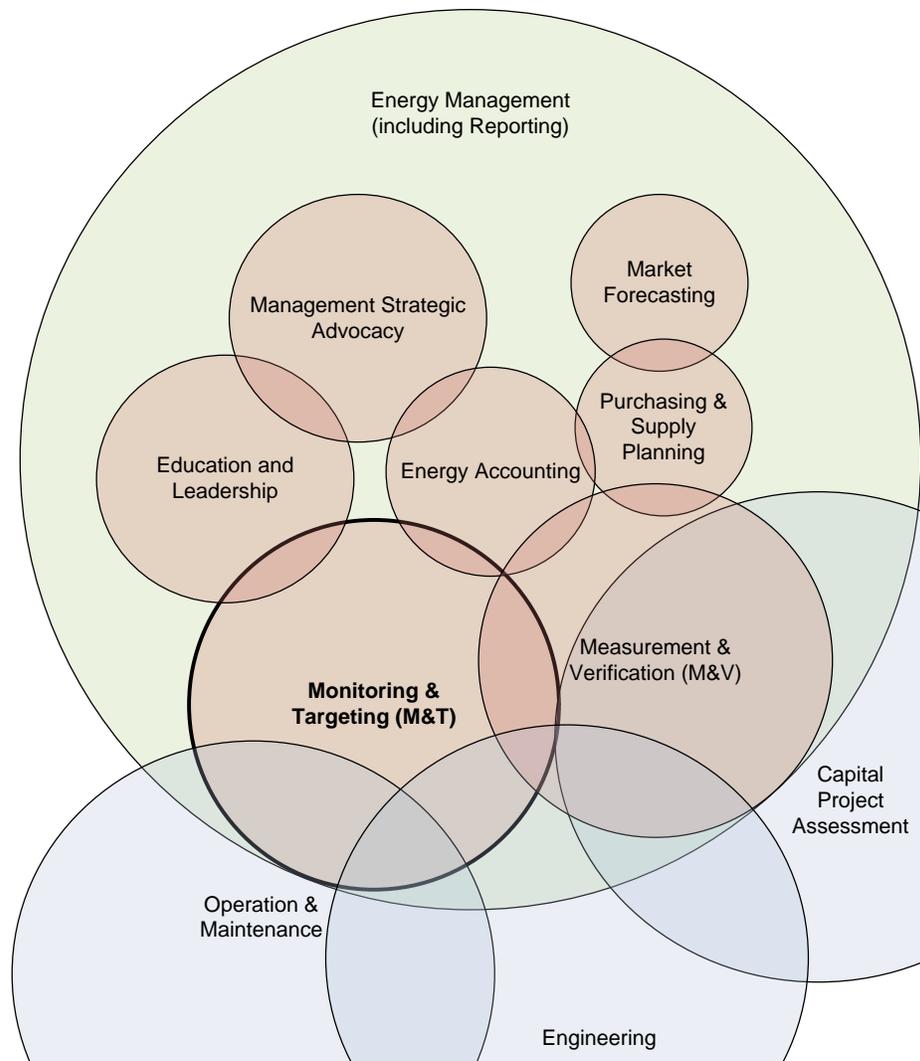
This dissertation investigates challenges in M&T in Chapter 2. To motivate the investigation, I briefly outline how M&T relates to energy management and why I chose to investigate it.

### 1.1.1. 'Work that EMs do', and differences between jobs

Monitoring energy is only a part of broader energy management (BRESCU, 2001). To put energy M&T in context, Figure 1 outlines some related tasks within (and beyond) energy management.

---

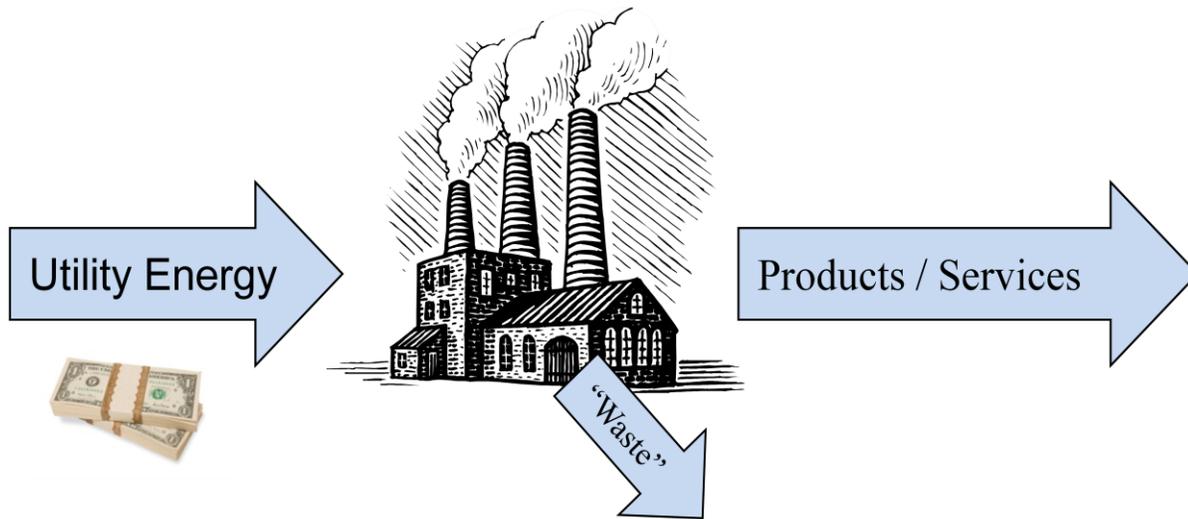
<sup>1</sup> One utility will subsidize more than 75% of an energy managers' salary for large customers (BC Hydro, 2010)



**Figure 1 - Conceptual Venn diagram of Monitoring & Targeting (M&T) and its overlap with other Energy Management and business operation activities**

Energy M&T is distinguished from other energy management tasks by being

- Business-facing, not market-facing. M&T examines energy consumption *within* a business, in the context of existing financial structures, not searching out cheaper prices.
- Operation and Maintenance-oriented. M&T aims to improve understanding of how energy is consumed, to inform better business operating decisions.
- Focused on discovering and distinguishing influences on energy use, more than quantifying accounting and contractual implications of known capital investments (as in Measurement & Verification).
- Integral with leadership and education. Except in very centralized businesses, Operation and Maintenance requires inducing effective work by colleagues.



**Figure 2 - M&T as identifying energy waste in a business**

Another way to express the objective of M&T is to “detect avoidable energy waste that might otherwise remain hidden” (Carbon Trust, 2008, p. 2), illustrated in Figure 2. Ideal M&T practice would rapidly and accurately identify all energy waste within a business, document unavoidable waste, and induce action on correctable waste (e.g. operation, maintenance, or capital investment), all while requiring little time or expertise.

### 1.1.2. Why focus on M&T for EMs?

Energy M&T is worth academic investigation for six reasons, in addition to the societal importance:

- It informs other energy management tasks. No control system can perform better than its measuring channel (Ashby, 1956). As energy management’s measuring channel, improvements in M&T should have potential to support other energy management work.
- It trades off human problem-solving labor for energy, substituting ‘neurons for electrons’, and is therefore an opportunity for human factors engineering.
- It is work that has been performed in very similar ways for 30 years (discussed in Chapter 2). Early texts on M&T give advice and identify challenges that continue to be identified in more contemporary writing, which suggests stagnation (CIPEC, 2010; Harris, 1989). If a field of practice has persisted but stagnated it suggests that current practices work well enough to have interesting structure but there may be un-realized opportunity for improvement.

- Industrial M&T is more sophisticated than energy conservation in the residential sector, which has been the focus of much research on efficiency behaviors (Abrahamse, Steg, Vlek, & Rothengatter, 2005). Human factors engineering practice is often to seek subject-matter-experts to observe and learn from, then design tools to help less expert practitioners achieve similar performance. There is potential to make more sophisticated M&T work accessible to novices.
- It is an opportunity for less-commercially-motivated academic research to contribute to the public good. Academics may be best positioned to investigate phenomena in M&T work since while M&T is locally profitable (International Energy Agency, 2014), it lacks the potential for order-of-magnitude returns on investment that attracts private R&D capital.
- Extensive background research did not find any accounts of M&T tasks having been studied in a controlled experimental environment.

Of course there are also pressing societal reasons to solve practical M&T challenges.

## 1.2 Societal Implications of M&T

The effects of fossil energy consumption-caused climate change on global ecosystems and human well-being continue to accumulate (IPCC Core Writing Team, 2014, p. 51). Climate change is most pressing global issue human civilization has faced (Stern, 2007). This provides a pressing societal motivation for understanding and overcoming barriers to M&T in industrial and commercial sectors which comprise 28% and 8% of global final energy consumption (International Energy Agency, 2012), more than the residential sector (23%). Climate change cannot be stopped with a silver bullet. “Silver buckshot” of emission-reducing interventions is required (Pacala, 2004). Energy management (and M&T) can influence the effectiveness of many energy efficiency measures, and whether measures are even adopted at all.

### 1.2.1. Efficiency and energy gaps

Energy efficiency at the end-consumer has multiplicative upstream benefits and is widely considered the best option for energy supply (International Energy Agency, 2014). Stabilizing global CO<sup>2</sup> emissions will require increasing global annual investment in energy efficiency by about 330 billion USD above current levels (IPCC Core Writing Team, 2014, p. 110). However energy efficiency measures, such as commissioning existing equipment (Mills, 2011) or adopting new technology (Jaffe & Stavins, 1994), are practiced less than economic analysis would predict,

leaving an “efficiency gap”. For example, foundries in Russia could reduce energy costs by 36% if they operated at the energy efficiency of their European equivalents (International Finance Corporation, 2011) This has been explained in economics terminology as due to imperfect information, irrational decision-making, or hidden costs such as transaction costs (Stern, 2007, p. 377). Even economists skeptical of an ‘efficiency gap’ acknowledge that “some consumers appear to be imperfectly informed, and the evidence suggests that investment inefficiencies do cause an increase in energy use” (Allcott & Greenstone, 2012, p. 5), and that “agents may be unaware of, imperfectly informed about, or inattentive to energy cost savings” (Allcott & Greenstone, 2012, p. 11). Competence at energy M&T is directly relevant to imperfect information and hidden costs in efficiency work.

### 1.2.2. Cognitive cost of Energy Efficiency

Hidden costs identified from economic analysis comprise capital and labor costs. Capital costs include information technology to reduce labor costs of configuration, monitoring, diagnosis, communication and action. In interview and field work, researchers have observed hidden costs as a barrier to energy efficiency adoption. A review of energy efficiency in the brewing and mechanical engineering sectors in Ireland, Germany and the UK found “the hidden cost that appeared by far the most important was the overhead costs of energy management, including the cost of employing skilled energy management staff” (Sorrell, Mallett, & Nye, 2011, p. 72). How effectively energy management staff develop insight influences whether subsequent efficiency work will deliver benefits. Data can be collected, but if it is not understood, it can contribute to “the lack of information about energy consumption patterns” (Schleich & Gruber, 2008, p. 1) found in the German commercial sector. Time is money, and while in heavy industries 5% energy savings means millions of dollars to re-invest in expert labour, for smaller businesses energy management work justifies little time, a leading barrier to effectiveness (Sorrell et al., 2011, p. 72).

### 1.2.3. Cognitive engineering to tilt equilibrium

The economic perspective on energy efficiency suggests a frame for cognitive engineering efforts: to reduce hidden (labour) costs and increase task effectiveness, hereby tilting the economic equilibrium of technology adoption and successful operation. Engineering interventions could include designing simpler, cheaper, or easier-to-use software tools that

improve work effectiveness. Work effectiveness would be improved by reducing expertise required, time required, likelihood of misses or mistakes, or other barriers to insight and understanding in existing energy management information technology. Opportunities are little-understood; “overhead costs of energy management ... do not seem to have been subject to serious academic study” (Sorrell et al., 2011, p. 25). However, opportunities are present:

*“solutions available for measurement, control, and improvement of manufacturing processes.... are not generally suitable for energy management in production on company, plant or process level. There remains a gap between the solutions available and the actual implementation in industrial companies”*  
(Bunse, Vodicka, Schönsleben, Brühlhart, & Ernst, 2011, p. 676)

M&T in energy management resembles other work that human factors engineering has aimed to support: investigation and diagnosis in technical systems. Human factors engineering theory and methods ought to apply to cognitive support for energy efficiency work (Flemming, Hilliard, & Jamieson, 2008; Moray, 1994; Vicente, 1998). This engineering dissertation contributes in three ways: through investigating M&T work challenges in Chapter 2, developing two prototype work support tools in Chapter 4, and explicating the design basis through a work analysis in Chapter 3.

## 1.3 Theoretical Work Analysis & Design challenges

This dissertation presents a Cognitive Work Analysis (CWA) of M&T in Chapter 3 that characterizes M&T problems in terms of a cognitive engineering theoretic framework (Bisantz & Burns, 2009; Rasmussen, Pejtersen, & Goodstein, 1994; Vicente, 1999). This analysis informed M&T tool design in Chapter 4, and provided a framework to develop an experimental evaluation of a novel M&T work support tool in Chapter 5.

### 1.3.1. Work Support Approaches

Cognitive work can be analyzed both to inform understanding of human cognition and to inform design. This dissertation applied CWA to inform engineering design. I observed M&T behaviors through field study in Chapter 2, but rather than seeking to explain their mechanism, I focused on how to overcome observed challenges. Methods for each differ since "the problem solving solution space for the scientists is a set of plausible cognitive theories, but the solution space for designers is a set of technologically feasible environments ... to obtain a given cognitive

solution" (Kirluk, 1995, p. 70). It was challenging to apply engineering analysis methods to M&T work for three reasons: domain variance, ecological problem-solving structure, and tractability.

### 1.3.2. Invariant constraints for formative design

The first challenge to analysis is that the technological environment of M&T varies widely between industries, which themselves undergo technological change. However, as discussed in Chapter 2, M&T methods have persisted with few changes from the 1980s to present. For an analysis to be worthwhile and contribute to the literature, it should capture constraints on energy M&T work that remain relatively stable across changing environments and technologies. The CWA engineering framework I used is intended to be applied for formative design (Vicente, 1999), but this requires determining which features of M&T work are most invariant and promising for information support.

### 1.3.3. Ecological problem-solving structures

Second, the M&T task can be performed many different ways that can be hard to compare. M&T is done with analytic, deliberate model-building strategies that are amenable to cognitive psychological approaches, and pragmatic, expertise-based strategies that focus on directly observing the environment to reduce the need for deliberation, more typical of ecological psychology approaches. At the same time, M&T is assisted by automated data collection and statistical analysis tools. A work analysis should ideally help compare and make design tradeoffs between cognition ‘in the head’, ‘in the world’, or ‘in the machine’ (Hutchins, 1996).

### 1.3.4. Tractable cost-effectiveness

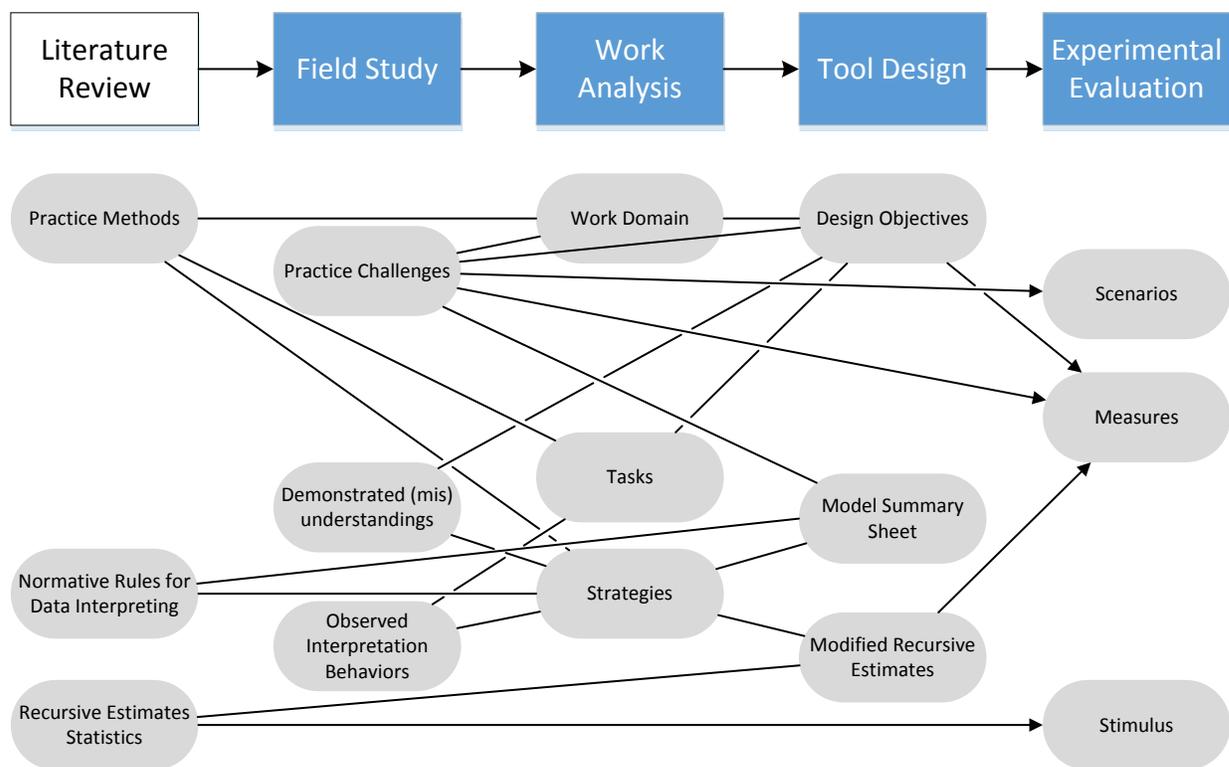
The third challenge was that the analysis needed to inform simple, cost-effective design appropriate for generally applicable M&T tools. CWA has often been applied to design bespoke interfaces tailored to particular complex systems (Vicente & Rasmussen, 1992, p. 595). Tailored M&T work support systems introduce more hidden costs, and I did not believe that such an approach would scale across industries, never mind have potential to scale down.

The analysis in Chapter 3 addressed these three challenges by analyzing M&T in terms of cognitive strategies that can be applied across many work environments and be used to flexibly exploit informative energy data or physical equipment structure. The most popular, statistical

model-supported M&T strategy was found to require rarely-communicated information about model structure for workers to resolve ambiguity in diagnosis. The analysis was applied informally to designs in Chapter 4, and to design the experimental evaluation of the work support tool in Chapter 5. The relationship between contributions in this dissertation will be outlined at the end of this introduction.

## 1.4 Structure and contributions of the dissertation

This dissertation is structured as a problem-driven, market discovery research program (Vicente, 2000), comprising four contributions: describing representative observations, defining design problems in a theoretic framework, developing two novel work support tools, and a controlled experimental evaluation of one work support tool. I discuss each of these four contributions in turn, with reference to Figure 3.



**Figure 3 - Outline of dissertation chapters and how findings from one section were applied in others. At top, chapters and associated contributions. At bottom, three aspects from each chapter.**

### 1.4.1. Contribution 1: Describing the M&T task in the context of Energy Management

The first contribution of this dissertation is the first reported field study of M&T work in Chapter 2. After orienting myself to M&T challenges through participant observation, I observed naturalistic industrial & institutional M&T behavior in industrial, hospital, and consultant work contexts. Field studies have a long tradition of supporting cognitive engineering work (Hutchins, 1996; McKay, 1992; Mumaw, Roth, Vicente, & Burns, 2000; Xiao, 1994). While the M&T pedagogical literature is very well-developed and work challenges can be inferred from instructional advice “very little academic research has been carried out in the area of energy consumption data analysis. A search of the literature reveals few papers published specifically on energy M&T” (Stuart, Fleming, Ferreira, & Harris, 2007, p. 1569). The overwhelming majority of academic research focuses on residential energy conservation behaviors (Abrahamse et al., 2005; Lutzenhiser, 1993), and design of residential energy control devices (e.g. home thermostats in Meier, Aragon, Peffer, Perry, & Pritoni, 2011).

This dissertation contributes to the literature by complementing the residential body of work with an industrial perspective (Hilliard & Jamieson, 2014b). Through studying a more complex work environment, I identified ineffective diagnosis as a barrier to effective M&T performance, and lack of information support for data and model assessment as a tool design opportunity.

### 1.4.2. Contribution 2: Characterize M&T task in CWA framework

The second contribution of this dissertation is to characterize the field study findings and literature review in cognitive engineering terms in Chapter 3. I applied the Cognitive Work Analysis (CWA) framework (Rasmussen et al., 1994) to describe the M&T information environment, tasks, and processing mechanisms in a descriptive and formative manner (Vicente, 1999). Some analyses have been published of other energy management tasks (e.g. energy auditing in Owens, 2013), but none of energy monitoring.

Task and work analyses provide a framework to help investigators and readers relate insights from other domains to energy M&T. The insights I identified from analyzing M&T work are:

- Developing and maintaining representations (energy-related data records and statistical models) are parallel tasks that M&T information support tools should minimize or support.

- Complex energy models used in some M&T strategies increase ambiguity and create trade-off costs of greater and more complex representation-maintaining work.
- Diverse M&T strategies enable diagnostic search through different structures of the M&T work environment. For example, a survey strategy can search through physical space, and sub-metering or equipment inventories can search through functional equipment.

I conclude that there is an opportunity for a diagnosis support aid that enables search of energy-related *processes* in a business system, while using no more complex data and models than those already used in standard M&T practices (Fawkes, 1988). This motivated a design effort in Chapter 4 to incrementally extend the simple statistical change *detection* strategy standard in M&T (Brown, Durbin, & Evans, 1975), to better support *diagnosis* through adapting a statistical strategy never before applied to energy M&T.

#### 1.4.3. Contribution 3: Apply CWA to M&T tool design

Chapter 4 describes the design of two inter-related software tools intended to support the information requirements of M&T diagnosis: a Model Summary Sheet, and Model Diagnosis Dashboard. The Model Summary Sheet describes the statistical basis for energy performance indicators in redundant formats accessible to the social groups involved in M&T. It explains what an energy model represents and the process by which it transforms independent variables (weather, productions) to estimate normative energy use. Existing M&T support systems generally conceal this information, presenting only the model predictions (Hilliard, Jamieson, & Jorjani, 2014).

The Model Diagnosis Dashboard builds on the Model Summary Sheet by enabling search through the structures represented in energy performance models using a modified Recursive Estimates (RE) change diagnosis strategy (Ploberger, Krämer, & Kontrus, 1989). I modified the algorithm and presentation format (Hilliard & Jamieson, 2013) using representation aiding principles (Vicente & Rasmussen, 1992; Woods, 1991) to make RE time-series charts behave as an analogical change indicator rather than a statistical certainty test (Hilliard & Jamieson, 2014a). This strategy-based approach to designing work support displays is novel in the CWA literature (McIlroy & Stanton, 2015, p. 154).

#### 1.4.4. Contribution 4: Controlled evaluation of M&T tool

The final contribution of this dissertation (Chapter 5) is the first empirical evaluation of M&T performance in a novel synthetic task. This controlled experiment recruited representative participants and evaluated their detection and diagnosis behavior in scenarios informed by the field study. Two experimental conditions compared performance using standard M&T tools with and without the RE diagnosis support aid described in Chapter 4. Results showed that diagnosis performance with standard M&T tools was indistinguishable from chance performance (28% correct) while access to the diagnosis aid significantly improved diagnosis success to 41%. Detection results corroborated accounts of M&T as a difficult task, showing a 57% false alarm rate across both experimental conditions.

The experiment structure could not distinguish whether the RE diagnosis aid better supported diagnosis of particular types of energy performance changes (which differ in the characteristic chart shapes they produce). However, the synthetic task, apparatus, method, and performance measures can be re-used for further investigation into M&T task performance.

### 1.5 Significance of the contributions

Academic study has contributed little to M&T practice, and little is known about how to reduce the hidden costs of M&T that hinder its adoption. This dissertation addresses this knowledge gap with a novel field study of M&T in industrial environments and by applying a cognitive engineering framework to model energy efficiency work. Compared to the early years of M&T practice, contemporary technology allows vastly more data to be collected and more sophisticated automation to control energy use. However, general-purpose tools have seen little innovation. This dissertation presents a novel application of a Recursive Estimates diagnosis support strategy to M&T. Finally, M&T task effectiveness has only been evaluated through practical trial and evaluation. This makes it difficult to objectively assess which aspects of M&T work people have difficulty with, or what tool support is effective. A controlled experimental evaluation of M&T is presented in this dissertation, which could be applied further to M&T education or tool development. The dissertation concludes in Chapter 6 by summarizing these contributions and outlining future research opportunities and development applications. Better information tools that support effective M&T diagnosis can speed energy management work and reduce a barrier to the drastic energy consumption changes necessary to address climate change.

## Chapter 2

### Describing M&T work as reported and observed

#### 2.1 Motivation: Comparing M&T norms with a descriptive account

This chapter describes Energy Monitoring and Targeting (M&T) in two ways. The first is a review of instructional and practice literature. Five approaches to M&T work are briefly summarized, one in detail: M&T through Recursive Cumulative Sum of Residuals (CUSUM) model-comparative analysis. The second description of M&T augments practice literature with a summary of findings from 1) an interview study, 2) participant observation, and finally 3) a field study of M&T practice. M&T field observations are worth reporting for two reasons:

- 1) Observing practice in a natural environment for a sustained period can attune researchers to invariant work challenges (Xiao, 1994), and
- 2) These work challenges can be contrasted with theoretic abstractions to identify issues that have been idealized out, as for example in contrasting algorithmic scheduling theory vs. professional practice (McKay, 1987, 1992).

Both justifications have obvious implications for designing more effective M&T work support tools. Ethnographic methods have a well-regarded and influential history in studying knowledge work (e.g. Hutchins, 1996; Zuboff, 1988). Field-based research is routine part of R&D, though no published accounts exist for industrial energy efficiency. Ironically, the human factors of energy efficiency have been almost exclusively studied in the simpler domain of residential energy conservation.

##### 2.1.1. Residential field observations informing policy and design

Like industrial energy efficiency, residential conservation was a topic of great interest following the 1970s oil shock. Governments sought to induce the general public to change their behavior (put on a sweater, turn down the heat) and make capital investments (purchase efficient appliances, renovate). Field studies were a key part of understanding how to induce these changes in residential energy efficiency. Ethnographers such as Willet Kempton studied how laypeople understood energy (Kempton & Montgomery, 1982), and how their understanding

influenced how they operated energy-intensive energetic processes such as home heating or cooling (Kempton, 1986; Kempton, Feuermann, & Mcgarity, 1992). A highly-cited finding from this work is a widespread naive mental model of thermostats as a ‘valve controlling heat flow’ (Kempton, 1986), an explanation of why many people turn thermostats all the way up expecting the furnace to heat the home faster. Home thermostats have remained a popular topic of field study, and a body of research has substantiated that the public largely do not use existing programmable thermostats as designers intended (Meier, Aragon, Hurwitz, Peffer, & Pritoni, 2010). Such evidence led to the US Environmental Protection Agency rescinding endorsement and subsidy of programmable home thermostats in 2009, instead developing a new program based on thermostat usability and behavior outcomes (U.S. Environmental Protection Agency, 2014).

Residential energy efficiency differs greatly from industrial energy efficiency. Home systems are smaller, homogenous (home HVAC, lights, appliances), and less energy-intensive. Motivating people to care is a barrier, and behavioral advertising campaigns, economic incentives are main design outcomes. Still, the extensive body of literature on residential energy conservation behavior (Abrahamse et al., 2005; Lutzenhiser, 1993) is illuminating, and in stark contrast to the lack of studies on how people behave when monitoring industrial energy efficiency.

## **2.1.2. Methods used to investigate M&T work**

Before introducing the literature review, I briefly describe the first two M&T field investigations that augment the third descriptive field study of Section 2.4.

### **2.1.2.1. Interviews with Industrial Energy Managers**

I first interviewed nine energy managers at large industrial companies and institutional facilities, and nine industrial energy efficiency consultants or utility program developers. Excerpts of the methods and findings (Hilliard, Jamieson, & White, 2009) that relate to M&T are summarized in this chapter. This interview study, like others (Sandberg & Söderström, 2003), was of limited validity in describing M&T. While all interview participants were involved somehow in industrial energy efficiency, only some were personally involved in monitoring or controlling business energy use. Similarly, this interview study was retrospective and did not include any task

observations. These limitations motivated me to gain personal experience using a second participatory investigation method.

#### 2.1.2.2. M&T Participant Observation

To understand first-hand the challenges of solving industrial energy efficiency problems, I conducted participant observation (Lofland, Snow, Anderson, & Lofland, 2006) by working as an energy M&T assistant in a steel mill from September-December 2010. Over four months, I worked to re-commission a disused M&T software system and re-introduce it to production managers. Later in a 2012 two-week consulting contract for a different company, I performed a less drastic model and interface refresh of the same M&T software. Together, the interview and participant observation helped me interpret M&T literature and plan the subsequent observational study that is the main contribution of this chapter (Section 2.4).

## 2.2 M&T Practice, past and present

The instructional literature on monitoring and targeting (M&T) energy use in industry is broad, and presents a variety of approaches to assessing energy efficiency. As outlined in Section 1.1.1 and the glossary, I use the term M&T more broadly than in the literature. I consider M&T the generic task of assessing and controlling business energy performance, part of energy management (EM) work. In both practitioner and academic literature, the M&T label specifically refers to a model-comparative statistical method, recursive CUSUM of residuals, which I review separately in detail in Section 2.3. Defining M&T more broadly serves as a reminder to consider complimentary approaches for achieving similar goals, which I explicitly describe in terms of cognitive strategies in Chapter 3.

Efficient energy use is obviously not a new concern, but as with many things, development is spurred during times of crisis. The seminal book on evaluating and improving industrial energy efficiency was written from British war-time experience operating steam systems (Lyle, 1947). The oil shock of the 1970s provided another crisis to motivate energy efficiency, and M&T practices were developed concurrently around the world, particularly in the UK and Japan, island petroleum-importing nations (Technological Economics Research Unit, 1979). Pilot applications proved potential effectiveness, educational materials were developed, and M&T practices have been promoted since. Several excellent guides to M&T have been developed by UK (Carbon

Trust, 2008; Gotel & Hale, 1989; Harris, 1989) and Canadian (CIPEC, 2010; Efficiency New Brunswick, 2010; Hooke et al., 2004) agencies.

The following sections review five approaches to M&T: management, model-driven monitoring, good housekeeping, real-time monitoring, and automated decision aids. These will be referred to later when describing field observations and performing work analysis.

### 2.2.1. Management Systems

Energy management (more broadly) influences whether M&T methods or tools will be effective. While this dissertation does not contribute to business management *per se*, I review management practices briefly since they determine tool-design-relevant factors such as:

- What M&T talent can be hired or retained,
- How much time workers will spend on M&T tasks compared to other priorities,
- What M&T findings are actionable (e.g. capital budgets, operation changes),
- How risk-tolerant the organization is to mistakes in diagnosis or corrective actions.

Four management applications that can be achieved using M&T methods and tools are:

- Operating accounting, (Carbon Trust, 2007; Fawkes, 1988; Gotel & Hale, 1989) assigning quarterly costs directly to business units to economically incentivize middle management
- Capital accounting, verifying energy returns on capital investment (ASHRAE Guideline Project Committee 14P, 2002; Efficiency Valuation Organization, 2012). I distinguish off-line, non-problem-oriented financial analysis as Measurement and Verification (M&V) rather than M&T (Figure 1).
- Management-by-exception, to focus scrutiny on under-performing business units or sites in a real estate portfolio (Carbon Trust, 2008; Gotel & Hale, 1989)
- Continuous improvement (Henze, 2001), such as the recent ISO 50001 (ISO Technical Committee 242, 2011)

These supervisory practices quantify energy consumption at financial scales (monthly by business unit), detect financially significant changes, and motivate employees (BRESU, 2001). Management tasks introduce overhead costs, and while they can secure supportive resources,

supervision is not sufficient in itself. Even if problems are detected, someone must still solve the work problem of localizing, diagnosing, and solving energy wastage.

### 2.2.2. Model-comparative monitoring

Model-comparative monitoring using CUSUM charts was the first method characterized as *Energy Monitoring and Targeting* (Aird, 1981; Fawkes, 1988; Gotel & Hale, 1989) and is still practiced today without any significant changes (Carbon Trust, 2008). It can be used to support accounting and management, but also front-line work. Because of its popularity and versatility, and because it forms the basis of the tools developed in Chapter 4, we discuss it in detail in Section 2.3.

### 2.2.3. Good Housekeeping

The model-comparative method just described was first formalized in the UK. However, Fawkes reports that UK industries generally used it for management accounting rather than as part of everyday operation and maintenance. Straightforward “good housekeeping” (Fawkes, 1986, p. 310) is to spot and fix obvious energy waste such as leaking compressed air, damaged insulation or idling equipment. Fawkes found that even in the 1980s good housekeeping was the norm in Japanese industry, according to a management philosophy of high employee participation in local problem-solving. These principles are a familiar part of the Toyota Production System.

Employee engagement and technology can substitute for each other. Fawkes notes "The absence of a high level of employee participation may explain the relative absence of successful good housekeeping campaigns in Britain and why many companies leap straight into the high cost technological route with all its costs and risks." (Fawkes, 1986, p. 313). Subsequent UK government instructional material urged that “a high priority should be given to measures in the [no-capital] category. Improvements in efficiency through better energy housekeeping can be substantial” (Gotel & Hale, 1989, p. 27). A weakness is that good housekeeping practices are difficult for management to supervise since they do not produce financially actionable records.

### 2.2.4. Real-time utility meter monitoring

Good housekeeping can be informed by utility meter data, indeed "Japanese companies stress monitoring and measurement above all else" (Fawkes, 1986, p. 5). I distinguish real time from

model-comparative monitoring as monitoring at finer frequencies (e.g. hourly), aggregation (e.g. by-building), or delay (e.g. available day-after) than the comparative model. If sub-meter data is interpreted directly it can be collected and interpreted in real time at whatever frequency (e.g. 5-minute) and aggregation (e.g. by-equipment) is effective and economic.

Benefits of real-time energy monitoring include the potential to reduce energy waste by taking corrective action quickly. Compared to other methods delayed by modeling frequency or data availability, energy meter data on its own can be provided soon after energy is used (Hooke et al., 2004, p. 29). This can potentially match the timescale and scope of M&T tools more closely with that of local work (Fawkes, 1986). The downside of real-time monitoring is “the paradox that in its sheer volume greater amounts of raw data become less easily intelligible as information” (Bobker, 2004, p. 173). More sub-meters mean more overhead costs and challenges of collecting, validating, navigating, and interpreting the resulting data. Without corresponding contextual data or understanding, workers may not be able to tell if metered energy consumption is good or bad. An advocate of real-time energy monitoring agrees that

*“unless this captured data is shared and analysed in an orderly and precise way that identifies problem areas and provides solutions, this mass of data is merely information overload. Data is not knowledge! Knowledge is information learned from patterns in data, and it follows that there must be the capacity and ability to convert information into knowledge in order to make sound energy-related business decisions” (Hooke et al., 2004, p. 2)*

Management must be committed to funding the capacity and ability of workers to act on meter data. However, information support tools design can potentially reduce the capacity or ability required.

### 2.2.5. Automated control or decision support

One way to derive value from large datasets is automated interpretation, such as decision support systems (Hooke et al., 2004, p. 13). Alarming extreme recorded consumption is straightforward, and more complex schemes are commercially available for heavy industrial equipment or process units that can be well-characterized and whose control may already be automated in a supervisory control scheme. If fault diagnosis rules can be specified and automated, decision support systems can “change the process performance reporting paradigm from ‘How did we

do?’ to ‘What prevented us from doing better?’” (Moore, 2005, p. 1). As with automation in other domains, automated M&T systems (or EMRS) can be attractive to engineering practitioners who believe “human factors contribute to poor performance in complex systems [and] rule-based EMRS systems circumvent these adverse human factors” (Moore, 2005, p. 4).

Costs and risks of automated diagnosis aids have been described in other domains. These include the costs of configuring and ‘teaching the machine’ (Sheridan, 2006), which for many systems may be substantial relative to energy costs, and desensitization to false alarms (Carbon Trust, 2008, p. 17). These costs increase for less well-characterized systems (e.g. job-shop manufacturing) where diagnosis rules are harder to pre-specify. However, the promise of lower labor means automated M&T systems will continue to be of commercial interest.

### 2.2.6. First principles energy efficiency analysis

This final category of M&T is engineering analysis often done by consultants or specialist technologists. This can range from an energy audit (CIPEC, 2010; Russell, 2009) to a first-principles re-configuration of a manufacturing process (CanMET Energy, 2003). This work is typically discrete (not continuous), off-line, and outputs both a technical recommendation and financial plan. This is borderline of what would be considered M&T, but being high-cost and high-complexity is a useful contrast to the energy monitoring practices discussed above.

### 2.2.7. Discussion: Tools for which approach?

The most effective way to perform M&T in a particular business will always be a combination of the above approaches. Each provides different benefits and makes different demands on executive management, front-line workers, capital, information technology, and skill. The variety of approaches to M&T is a challenge to software design. “There is no ‘one size fits all’ approach to the issue. As the purpose of installing an EMIS is to provide information to people that enables energy improvement actions, the organizational context that drives those actions is paramount to EMIS success. In other words, an EMIS alone will not save money.” (Efficiency New Brunswick, 2010, p. 11). Features raise software cost, but before starting M&T, the customer may not know what features they need. Managers may settle for straightforward business accounting approaches, while engineers may desire a complex automated system. Early on, M&T practitioners found “experience has shown that it is often better to make a start with a

simple system and achieve an early success. An unnecessarily complex system should be avoided" (Gotel & Hale, 1989, p. 8).

An ideal M&T work support tool would support as many approaches to M&T as possible, enable success at front-line energy conservation work, and be simple to maintain. Model-comparative monitoring with CUSUM charts is one of the most popular existing approaches, possibly because it best satisfies these criteria.

## 2.3 M&T with Recursive Cumulative Sum of Residuals (CUSUM)

This section presents M&T by CUSUM charts, since it was the M&T method observed most frequently in the field study (Section 2.5) and was the basis for a novel M&T tool design in Chapter 4. History, application steps, interpretation, and challenges are presented in turn.

### 2.3.1. Origin of CUSUM method

The Recursive Cumulative Sum of Residuals (CUSUM) algorithm is one of the first methods suggested for statistically detecting changes in engineered processes. It was first introduced as part of monitoring steam plant efficiency during the Second World War (Lyle, 1947), based on similar principle as Shewhart's Control Chart of 1924. CUSUM charts plot integrated residuals (rather than raw variance), which is appropriate since the costs of energy waste are proportional to sum total, not variation. CUSUM was more thoroughly investigated for general econometric use with a formal test statistic (Brown et al., 1975; Page, 1961) and extended by substituting a parametric model prediction to adjust for environmental or business conditions. CUSUM was first proposed for M&T in UK technical reports (Aird, 1981; Technological Economics Research Unit, 1979), then described in comprehensive guidelines (Gotel & Hale, 1989; Harris, 1989). The benefits of adapting a Control Chart-based approach were attractive to engineers:

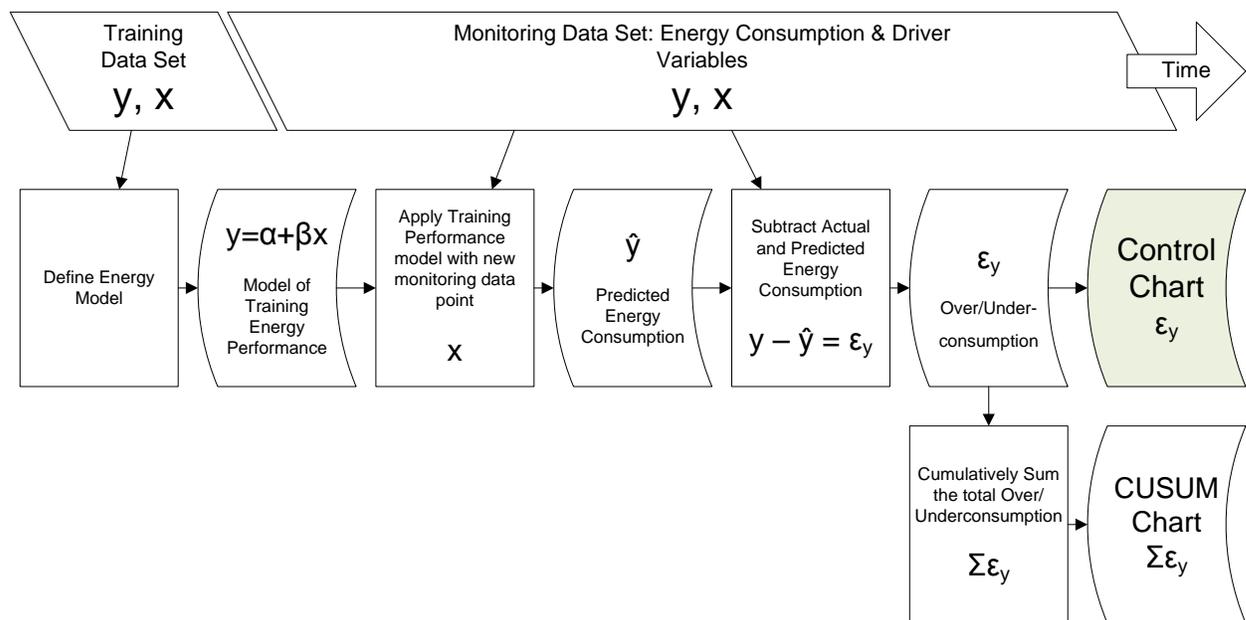
*"There is an important parallel between money management and energy management. Information in a business usually flows in the form of what is called time series information; that is, information which accrues as time goes by, organized in time order. However, the most fundamental indicators of business performance are not determined by analyzing financial information whilst it is in this form. The problem is then how to analyze it differently; so that it brings out the fundamentals which affect the performance of the business, but retaining enough of the time series element to enable other people in the enterprise to relate*

*it to their day to day activities. The techniques for doing this have been long established for managing money. They rely on the economic laws of supply and demand and the time value of money. For energy, they simply depend on a different set of laws - the physical laws of heat and work." (Harris, 1989, p. 7)*

CUSUM remains the most commonly suggested statistical aid for measuring and verifying (M&V) energy savings (ASHRAE Guideline Project Committee 14P, 2002) and for performing M&T (Carbon Trust, 2008; CIPEC, 2010; Hooke et al., 2004).

### 2.3.2. Method Outline

The core steps of applying CUSUM charts to M&T are illustrated in Figure 4. They are:



**Figure 4 - Data transformation steps in CUSUM algorithm. Consumption (y) and driver variables (x) at top, processed time-series charts at right.**

- 1) Train a model to describe business energy consumption  $\mathbf{y}$  as a function of some measured independent variable(s)  $\mathbf{X}$  (i.e. energy driver(s)). Linear regressions based on historic data are common.
- 2) As new data is collected, apply the model to the drivers and to calculate model-predicted consumption  $\hat{\mathbf{y}}$ . Subtract this from actual consumption to calculate the residual error  $\boldsymbol{\epsilon}_y$ . This over / underconsumption time-series is often plotted as a Control Chart line graph.
- 3) Integrate the residuals, and plot this as a second line graph, the CUmlative SUM chart.

- 4) Inspect the consumption and CUSUM charts. Interpret inflection points, where the slope of the CUSUM chart changes, as times where some business energy performance change may have occurred. Investigate.

Of this four step process, model-building and chart-interpreting present the most challenges, and are the focus of the contributions described in Chapter 4.

### 2.3.3. Model Building

Developing a statistical model is familiar to empirical researchers and explained in many M&T guides (Carbon Trust, 2008; CIPEC, 2010; Harris, 1989). We discuss it briefly here in the context of M&T practice, and to contextualize field study observations. First, and obviously, the CUSUM method requires data at least from utility (e.g. electric or gas) meters, and ideally independent data associated with energy use (discussed below). Data will be collected at some frequency, aggregation, and processing delay. Data availability and quality define which models can be built and therefore what CUSUM charts can indicate.

Models define the meaning of CUSUM chart slope, their most salient visual feature (Figure 5). An upward CUSUM slope represents consumption greater than model-predicted and vice-versa<sup>2</sup>. Models can be developed according to standards (ASHRAE Guideline Project Committee 14P, 2002), but regardless model meaning can vary greatly depending on how they are formulated. From simplest to more complex, models can describe energy performance as:

- An arbitrary or carefully chosen value (e.g. industry benchmarking)
- A mean historic consumption
- An empirical model fit to historic consumption and recorded conditions
- A first-principles model in terms of ‘driver’ data developed by engineering analysis

Ordinary least-squares regression on historic performance was first proposed (Harris, 1989), though in principle any modeling method can be used such as piecewise mean historic averages (Stuart et al., 2007). This dissertation will focus on regression models, as they were the only

---

<sup>2</sup> Though this is by convention and opinions differ whether ‘up is good’ or ‘up is overconsumption’ is more persuasive or memorable. The convention is difficult to remember, as shown by a typographic error corrected after publication in a seminal work (Harris, 1989, p. 51).

approach seen in participant observation and the field study described below. Regression models require data variables that:

- Are correlated with (and ideally cause) energy consumption. M&T practitioners call these independent variables “energy drivers” (Fawkes, 1988).
- Have little time lag between driver ‘cause’ and energy ‘effect’. Such delays define the minimum frequency at which regression models are useful for CUSUM charts<sup>3</sup>. Daily frequency is usually the limit for most applications, e.g. to building heating, and in systems with energy storage or long time lags weekly frequencies may be all that is possible.
- Can be associated with energy using linear parameters. Variables can be transformed to suit linear models such as by scaling, e.g.  $\sqrt{\text{Wind speed}}$  or piecewise break-point conversion, e.g. from outdoor temperature to Heating Degree Days (HDD) (Aird, 1981; Fawkes, 1988).
- Are (in my experience) ideally measures of extensive properties (e.g. mass, volume, energy), which sum over space and time. Drivers that measure intensive properties (e.g. temperature, pressure) create models specific to the trained aggregation and frequency. Intensive data limits how models can be applied, preventing use with other data only available at longer timescales (e.g. monthly utility bills).

Finally, crucially, and subjectively, M&T models require training data over a training period that represents meaningful, representative business behavior. Practitioners recommend the “longest and earliest consistent” (Harris, 1989, p. 49) period that fairly weights all conditions or modes of operation of the business (ASHRAE Guideline Project Committee 14P, 2002). A full year of data is not an uncommon requirement, which in the field study was observed to have practice implications (Section 2.5.5).

### 2.3.4. Model Application and Calculations

CUSUM charts are developed by applying the model to incoming driver data to calculate *energy performance at model-trained system condition*. The process of applying energy  $y$  and driver variables  $X$  is shown in the flowchart of model generation and application in Figure 4.

---

<sup>3</sup> Autocorrelated models are possible but introduce other problems. A practical solution to time lags is to reduce the modeling frequency (e.g. from daily to weekly intervals), or use a more sophisticated modeling method.

Obviously CUSUM methods cannot be applied to time periods where driver data is missing. Delay in data measurement and collection determine how retrospective CUSUM charts are (Hooke et al., 2004, p. 29). For Control Chart use, each missing data point is independent and does not taint adjacent time series points<sup>4</sup>. However since CUSUM is an integral, uncertainties from missing data propagate into the future and alter the meaning of the CUSUM time-series as *summed total energy over/underconsumption*. Data quality must be maintained for CUSUM charts to be valid.

### 2.3.5. Interpretation

The popularity of CUSUM charts has been attributed to their “distinct advantages in M&T procedures” (Gotel & Hale, 1989, p. 22). These advantages are not just numerical but also perceptual. While the same data can be presented as Control Charts, small changes in mean energy performance residual may be difficult to perceive on a varying scatter- or line-plot. By integrating the residuals, CUSUM charts transform a change-in-position perceptual problem to a change-in-slope perception task. Additionally, the y-position of a CUSUM chart can be interpreted as *cumulative savings/loss over time period X*, with obvious financial applications.

CUSUM’s integrated residuals create a line chart with perceptual features whose interpretation can be taught in instructional pamphlets (Carbon Trust, 2008). These are:

- Horizontal CUSUM chart segments represent the same energy performance as captured in the model training period
- Straight line segments in CUSUM charts represent time periods of constant energy performance
- Steps or sharp jumps in the CUSUM plot correspond to brief times of changed energy performance (or more likely, instances of bad/missing data).
- Slope changes (inflection points) signify when a persistent change may have occurred

However, these simple rules are not entirely reliable.

### 2.3.6. Problems with CUSUM Interpretation

CUSUM charts can be ambiguous, ineffective, or misleading under certain circumstances. Interestingly, the M&T literature of the 1980s discusses these weaknesses at greater length than

---

<sup>4</sup> Except for models with autocorrelation terms.

more contemporary material, possibly because the then-novelty of the CUSUM method warranted reflection. Difficulties found in the literature are discussed below, while those I observed from participant observation or the field study are discussed in Section 2.5.

CUSUM plot scales can be confusing to interpret. To start, CUSUM is in energy units which may not be meaningful to an M&T analyst's colleagues. Secondly, chart shape can be misleading since zoomed-in CUSUM chart excerpts can be self-similar to larger portions. Interpreting the magnitude of energy savings/losses requires comparing to average daily consumption or cost in dollars. Early graph-paper-based guides recommended using transparent CUSUM protractor overlays to provide reference slopes corresponding to savings percentages (Gotel & Hale, 1989; Harris, 1989). More recent practitioners may have assumed that graphing software has overcome the need for visual context.

A third difficulty is that CUSUM slope changes that are gradual, curvy, or periodic are less perceptible (compared to sharp edges) (Appelle, 1972) and more ambiguous. Harris (1989) discusses characteristic CUSUM chart shapes he terms “scalloping” (Harris, 1989, p. 23) caused by seasonal changes in winter heating or summer cooling of buildings. Harris suggests resolving ambiguity with other representations (e.g. scatterplots) or comparing against driver data.



### CUSUM Ambiguity: “Scalloping”

**Figure 5 - CUSUM chart illustrating how a single business change (at top center on 1 Feb, 11) can produce an wavy, ambiguous CUSUM chart. Overlaid straight lines show how standard CUSUM interpretation rules might be (mis)applied. Such changes can be associated with an intermittent operation mode (e.g. excess fresh-air ventilation), or a true permanent change (e.g. a hole in a wall or window).**

Fourth, changes in energy performance that occur only in particular operating modes or environmental conditions are not always distinctly represented in a CUSUM plot, which makes the chart less diagnostic. Finally, CUSUM charts' response to driver or consumption data that is

missing, mis-measured, or represents non-meaningful conditions may not be distinguishable from true changes in the system. Literature recommends that:

*"It is essential to ensure that variations in input data are not caused by faulty measuring equipment or sensors. ... discount abnormal months ... abnormal consumption patterns must be filtered out" (Carbon Trust, 2008, p. 5)*

The information that workers need to judge ‘abnormal’ consumption patterns are not further explored, though an automated statistical approach to data validation is an option (Capehart & Capehart, 2005, p. 440).

### 2.3.7. CUSUM Conclusions

This section introduced principles, method, and caveats of performing M&T with CUSUM. It is one of the most common M&T methods and is widely recommended in the literature. While some caveats of practice are mentioned in instructional texts, it is not clear whether they adequately describe how M&T and CUSUM chart-based tools are applied in practice. The next section summarizes a field study conducted to augment the literature review.

## 2.4 Observational Study

To augment the interview and participatory studies described in Section 2.1.2, and discover how practitioners performed M&T using methods such as CUSUM analysis, I conducted a field study in the fall of 2011 over five months in two types of naturalistic setting. The field study was conducted in partnership with energy analytics company Energent Inc., whose employees and clients consented to participate. Field study methods and findings are discussed more thoroughly in other publications (Hilliard & Jamieson, 2014b). We briefly introduce the study methods here, then outline findings relevant to 1) the CUSUM chart discussion above, 2) the subsequent work analysis, and 3) opportunities for M&T tool design improvement.

### 2.4.1. Environments Observed

The environment in which work takes place should influence how experts will perform M&T interpretation (Bröder, 2003; Payne, Bettman, & Johnson, 1993; Rasmussen et al., 1994). Factors the literature suggested would influence strategies were:

- Effort to access data sources (such as ability to directly observe business activities or ability to use cumbersome software),
- Accuracy and richness of data sources (such as being able to access finer timescale data, or primary reports of events),
- Time pressure or workload (particularly for strategies that require on-going maintenance)
- Opportunities for action (as experts sample enough data to decide between anticipated courses of action) (Rasmussen et al., 1994)

To sample across these factors, the field study was planned to sample from two categories of M&T work environments:

- **Off-site at an energy service company office** where statistically competent analysts could indirectly access clients' shared data history, with less time pressure, but could not act directly.
- **On-site in an office environment** where energy 'specialists' had indirect experience with energy-consuming work, access to machine-readable data history, could question colleagues, were under some time pressure, and could act directly or collaboratively.

The study planned but failed to sample from a third environment, **On-site in an operations environment** where operations staff had direct experience with energy-consuming work, could access non-machine-readable records and activities, might be under urgent time pressure, and could act directly on equipment.

#### 2.4.2. Workers

Workers with different aptitudes will be capable of performing M&T in different ways and will have different preferences. Some may:

- have only learned a few strategies (e.g. automated agents, novices),
- be more effective at using certain strategies (e.g. have aptitudes, support tools, or practiced expert perception / cue attention),
- have differing 'subjective task formulations' (e.g. engineers want to 'solve the mystery', technicians want to 'fix it fast') (Rasmussen & Jensen, 1973), or
- be willing to expend more or less effort and demand more or less accuracy (Payne et al., 1993)

Therefore the study recruited participants from each of the environments (Table 1):

**Two Energy Analysts from the energy analytics service company** with detailed M&T-related education but no access to non-machine-readable day-to-day data, and only a coarse understanding of the clients' businesses.

**Two Energy specialists** responsible for energy at client businesses, with some M&T training, a good understanding of overall energy use patterns, and a moderate understanding of the work system.

**One Operations specialist** at a client business who was expected to have limited M&T training, but a detailed understanding of work practices.

**Table 1 - Participating institutions and workers in observational field study**

<i>Site</i>	<b>Approximate yearly utility consumption</b>	<b>Participants</b>
<i>Client Site A: Chemical Manufacturer</i>	5 GWh electricity 8 Mm <sup>3</sup> natural gas ~\$1.3 million	Energy Specialist #1
<i>Client Site B: Hospital</i>	25 GWh electricity 4Mm <sup>3</sup> natural gas ~\$3 million	Energy Specialist #2 Operations Specialist #2
<i>Energy Management Information System Supplier</i>		Energy Analyst #1 Energy Analyst #2

### 2.4.3. Methods

Methods are described briefly here and in other publications (Hilliard & Jamieson, 2014b). The five participants were observed at work over eight weekly sessions between October 11<sup>th</sup> and December 9<sup>th</sup> 2011. In each session they were asked to 'think aloud' (Chi, 1997; Ericsson & Simon, 1992) as they inspected recent business utility consumption. Participants were interviewed twice, once at the beginning of the study and once at its conclusion, with questions addressing their motivations, experiences, and understanding. Recordings and field notes were interpreted qualitatively (Sanderson & Fisher, 1994). I attempted to code behavior fragments to perform a quantitative analysis, but had difficulty achieving inter-rater reliability and did not further pursue the approach. Results from the field study showed behavior that contrasted with

instructional M&T literature, which informed the subsequent Cognitive Work Analysis in Chapter 3.

#### 2.4.4. Limitations

This field study has several limitations, primarily that it did not systematically sample M&T practice. The three participating organizations (factory, hospital, and energy information system provider) were recruited through a convenience sample. This sampling limitation also restricted the field study to only one M&T software product (Energent's). The four energy specialist and analyst participants were practiced, but not expert, and the fifth a complete novice.

The resulting behaviors are not completely representative of M&T practice. Heavy industrials with ten times greater utility bills would be expected to support more expert behavior. However, from literature review (Section 1.1), and earlier interview and participant observation studies (Section 2.1.2), the sophistication of M&T even at large industrials varies greatly. Furthermore, well-designed M&T work support tools should enable even non-experts such as those observed to be effective. Therefore the difficulties observed in M&T practice by non-experts in a Canadian medium enterprise context should still be useful to inform design.

### 2.5 Results of M&T Observations

The observations described below are from my perspective as an outside observer, with an engineering graduate student's perspective. Ethnographic personas describing participants' motivation and perspectives are presented elsewhere (Hilliard & Jamieson, 2014b). The findings informed subsequent analysis and design activities.

#### 2.5.1. Participant discoveries and learning

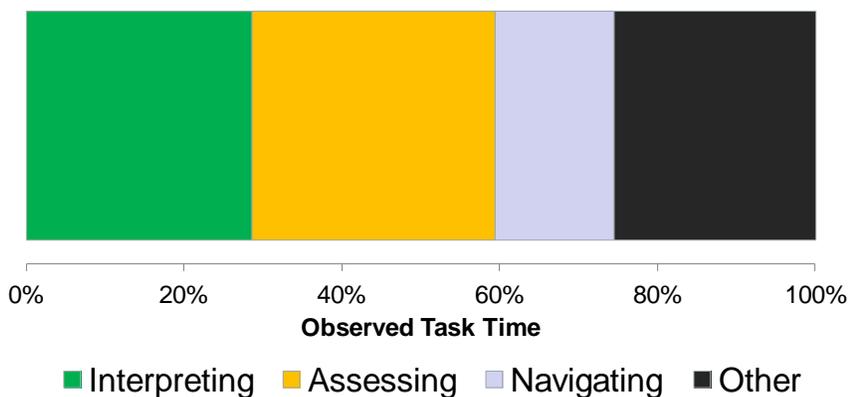
A summary observation is that in the 8-week observational study, none of the 5 participants discovered nor corrected any specific energy-wasting operation practice or maintenance issue. Participants' limited task success is a limitation of the field study, but is consistent with M&T's reputation as a difficult task to perform cost-effectively. On-site client participants instead sought confirmation of the effects of known changes. Participants at the hospital remarked several times on the clear over-consumption effects of their newly-constructed building wing and the operational impact of an electric transformer failure. The manufacturing industry's energy

specialist used the M&T tools to confirm the energy consumption effects of a known site shut-down and the production team's switch to a 5-day schedule.

Participants did report gaining a better understanding of how to use the M&T tools, and both energy specialists learned something about how their site consumed utility energy. The factory specialist learned a rough estimation of the consumption of different process units (discussed below in terms of mental models), and the hospital specialist estimated the functional capacity of the heating system (discussed below in terms of a thought experiment).

### 2.5.2. Data Records

Participants spent more time assessing data records at both sites than I had expected based on my participant observation. While I spent most of my M&T participant observation (Section 2.1.2.2) assessing data and developing models, I hoped this would not be needed in established practice with a commercial product. By *assessing* I mean seeking to better understand the quality of data or models in representing phenomena, with the goal of determining whether to believe them or intervene to repair the data record. While my classification of participant behavior is unvalidated (see Section 0), I estimate that participants spent roughly 1/3 of their task time assessing data or models (Figure 6). Of time spent assessing, roughly 2/3 was considering M&T model quality and 1/3 utility and driver data.



**Figure 6 - Rough proportion of M&T task time analysts and energy specialists spent interpreting energy data, assessing data or models, and navigating the M&T information system. Coding not validated.**

Specifically, the factory had issues with delayed data collection. Retroactive production record-keeping did not always get updated in the M&T system. If recent data had false 'zero' entries,

the M&T model would under-predict. At the hospital, less data was collected (only utility consumption, outdoor temperature and day-of-week), yet complete data was still delayed by up to 24h.

Manual inspection of data tables or charts was the only way the M&T tool supported assessing data. Gaps in meter data or order-of-magnitude errors were recognizable on CUSUM charts, but zeroed driver data or mis-calibration were not. In post-study interviews, only the factory energy specialist from Site A accurately described the steps involved in data measurement, processing, and transmission, with the opportunities for distortion. One shortcut approach used by an energy analyst to assess both data and models was consistency. If the M&T model was “in the ballpark” of actual consumption, he considered data and models likely all right. If the model and utility data disagreed, he had to investigate both to diagnose which might be mis-representing.

### 2.5.3. Energy Interpreting

Participants most often interpreted energy consumption (~1/3 of the time) with respect to an energy performance model of their site, either explicitly or implicitly with CUSUM charts. Besides interpreting time-series CUSUM charts (discussed below), participants also inspected utility meter data alone, sometimes plotted against energy-related data (e.g. outdoor temperature) on the same chart. Participants’ next most frequent context sources to interpret energy were system structure (controls or equipment functions), historic energy consumption, measurable disturbances (e.g. weather), and temporal reference systems (e.g. day-of-week). I rarely observed participants referring to colleagues’ suggestions, unit conversion, or abstract energy theory.

Participants’ use of context to interpret energy data was likely influenced by Energent’s M&T tool. Model output and CUSUM chart were charted by default, and the behavior of the tool may have discouraged use of other comparators. For example, the tool only reported units of binned energy per time step (e.g. kWh). However, energy units are not invariant across timescales. On two occasions when an analyst 'zoomed in' to inspect the factory meter data at finer timescales, they became disoriented when they could not find the numeric value of the feature they were investigating (e.g. because 24 kWh / day became 1 kWh/h).

By contrast, units of average power (e.g. 1 kW) are invariant across time scales, which should better support expertise in developing memorized ‘lookup tables’ (Dutton & Starbuck, 1971;

Rasmussen, 1986) to interpret system behavior. Power units were also described by the hospital's analyst as being useful to compare with knowledge of system structure. When discussing with on-site workers, for example, analysts converted energy to trade units to help prompt recognition of suspect equipment (e.g. 2.2 horsepower for an idling motor, \_\_\_ million BTU/h for boilers).

#### 2.5.4. Understanding of business structure

Business system structure was frequently mentioned as context for making sense of energy data, distinct from statistical models which themselves implicitly represent structure. By business system structure I mean physical layout, equipment, constraints on operation, and other invariant properties that affect business energy consumption.

Participants expressed some difficulty in understanding system structure. The hospital energy specialist described heating processes with what seemed to be folk models (Kempton, 1986), saying the building consumed energy "like a sponge", that "it is hard to get it hot, but once it is hot it is nice". This understanding explains the time lags needed to prepare heating equipment before occupancy, but does not suggest any benefits from not climate-controlling unoccupied space. Such misunderstandings may correct over time – the hospital's energy specialist intended to use M&T to support trial-and-error experimentation in operations strategies, which would reveal effective actions regardless of incorrect mental models.

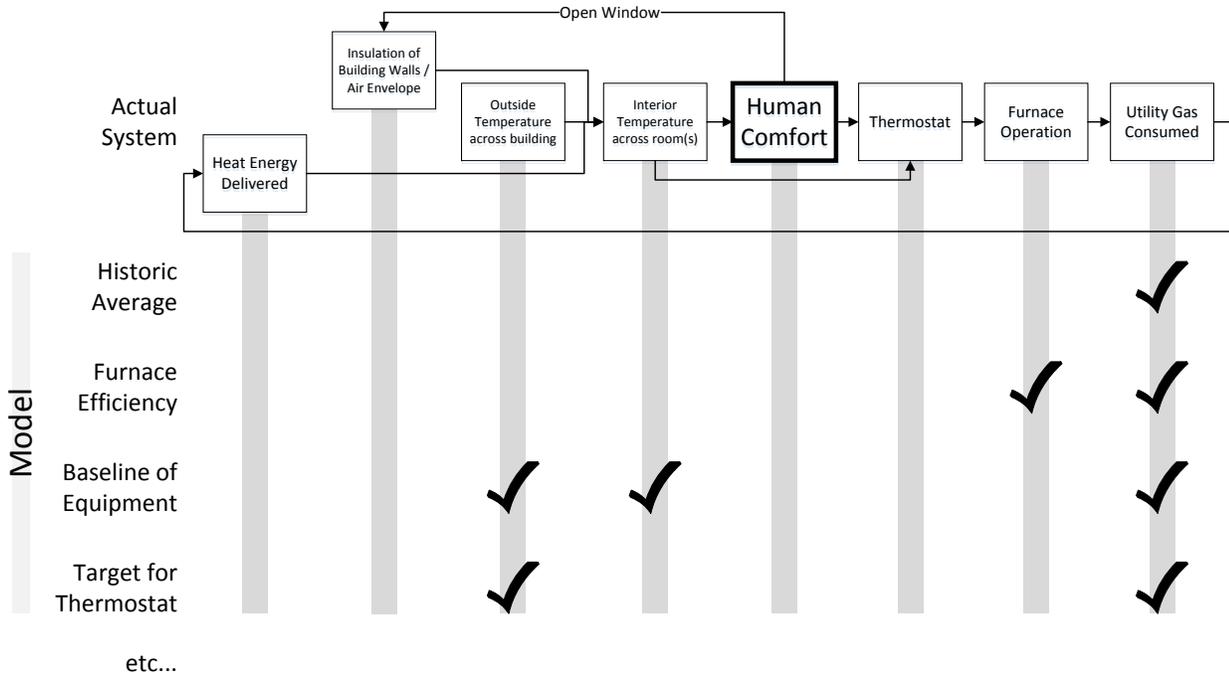
Participants also expressed difficulty keeping their understanding of business structure up-to-date. At the hospital, the energy specialist complained about how hard it is to learn and remember site equipment and controls for a huge facility with many contractor designers. He relied on binders of documentation to answer questions such as "which air damper does what"? Energy analysts were even less aware and in interviews could not name specifics of equipment installed at either site. Analysts referred to stereotypical types of M&T clients (e.g. "this is a hospital", "this is a manufacturing plant") to judge the plausibility of energy consumption records. Similarly, they referred to energy model structure to infer which types of equipment might be installed at a site ("this site probably has a chiller"). Information about specifics of site structure was not recorded anywhere in the M&T tool.

### 2.5.5. Model Developing

The M&T tool allowed selecting between many pre-developed linear regression energy models (Section 2.3.3) for each site. When the study began, the models in use had been created by analysts' predecessors and participating analysts were not familiar with them. Models seemed to serve multiple purposes, some anticipated by tool designers, some not. These included serving as a:

- “Baseline” (Gotel & Hale, 1989) to measure and verify (M&V) energy savings for contractual payments (ASHRAE Guideline Project Committee 14P, 2002)
- “Target” to motivate behavior changes from co-workers (Gotel & Hale, 1989)
- Diagnosis aid to interpret recent energy use (for problem solving)
- Idealized reference for business energy structure (and behavior), as just mentioned (Section 2.5.4)
- Encouragement tool to substantiate management success stories

This (incomplete) list of model purposes implies different choices in model-building. For example, a “Target” motivational model would *omit* explanatory variables that the designer is trying to encourage workers to control.



**Figure 7 - Example of purposeful model design for a heating system (as a block diagram, top). Models (at left) could include (checkmark) measures of certain processes to isolate furnace combustion and/or building envelope insulation. A simpler model (bottom) might omit measures of indoor temperature so that the resulting CUSUM chart would respond to turning down the thermostat.**

An example of purposeful model design choices is whether to include thermostat setting in a model of heating efficiency (Figure 7). Another example is whether to model manufacturing energy consumption using measures of total production or "good" production (minus scrap). Either can be justified depending on model purpose. A model using Total production:

- Could be applied to infer manufacturing equipment performance
- Could compliment a production team already motivated to reduce scrap

By contrast, a model using "Good" production (without scrap):

- Reflects the net effect of any tradeoffs made between energy and quality control (e.g. benefits of over-heating steel to leave a processing time buffer before cooling and recrystallizing)
- Would follow the principle of 'models should compensate only for uncontrollable factors'.

Different model purposes were also more vulnerable to known, un-actionable changes in system structure. At the hospital, a renovation was in-progress and an electric transformer had failed (requiring large equipment to idle rather than be started on demand). At the factory, production schedules changed from a 7-day to 5-day operation plan. These changes had occurred since the

most recent energy performance model had been calibrated. “Baseline” models are intended to highlight and quantify such changes. But for models to serve as a “Target” or problem-solving aid, these known structural changes required users to estimate whether overconsumption was more than expected for the known changes. This created interpretation problems, discussed in the next section.

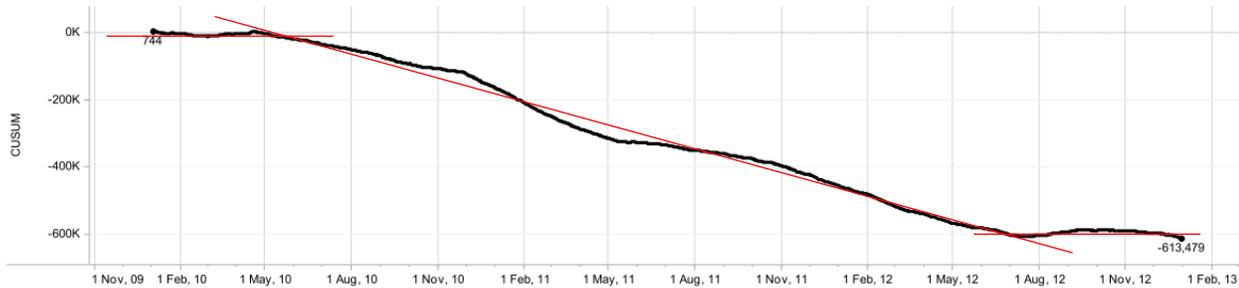
An obvious solution is to re-calibrate the model to account for known structural changes. However this requires waiting long enough to collect a dataset that fairly represents system performance, ideally a full year of operation (ASHRAE Guideline Project Committee 14P, 2002) but at least several months. In the meantime, over the first month and a half of the study, participants found energy models less informative or useful as diagnosis aids. These difficulties were compounded by how little information about energy models the M&T tool made accessible to users. For example, for the first four weeks of the study, even the energy analyst assigned to Site A did not notice the limitations of a Site B model that was old and trained on inadequate data. Later in the study after analysts revised electricity and gas models for each site (to reflect the above changes to system structure), client energy specialists were not notified. On both occasions for the hospital the energy specialist did not notice that the CUSUM chart appearance had changed drastically from the previous week. These challenges in understanding model meaning motivated one of the redesign concepts pursued in Chapter 4.

### 2.5.6. CUSUM Interpreting

Three recurrent difficulties in interpreting CUSUM charts were observed in addition to that described in the instructional literature (the ‘scalping’ CUSUM charts of Section 2.3.6). These included a) Large changes obscuring later changes (as just discussed in 2.5.5), b) Ambiguity in perceiving overlapping (small) changes, and c) Inconsistency in event-based explanations.

Large known structural changes (as discussed above) transform a CUSUM chart from a generally horizontal plot to a diagonal line, which obscures subsequent, smaller energy performance changes in two ways. First, it expands the vertical axis and compresses the chart scale (See Figure 8). This means that jumps or other transient patterns will become difficult to perceive. Second, instead of a horizontal line with which to compare angled lines, the chart becomes a set of obliquely angled line segments, which are less distinguishable (Appelle, 1972). In such

situations the Control Chart may be more useful since it shows deviations as a vertical shift in mean, rather than a change in slope.



### CUSUM Scale Inconsistency: “Stale Model”

**Figure 8 - Example of a CUSUM chart illustrating how a large system change (top left) creates a diagonal feature (center) that obscures smaller changes that would be more perceptible otherwise**

The second challenge observed is related to the known issue of CUSUM “scalping” (Harris, 1989, p. 23) due to changes in driver sensitivity (Section 2.3.5). When such driver-related changes overlap, even if changes do not obscure each other (as in Figure 8), they produce ambiguous CUSUM plots (Figure 9). For example, if consumption is shifted from one period to another (e.g. weekday to weekend), or if a driver-related efficiency change is offset by a coincident change to baseload, a CUSUM chart will show only the net difference. This difference will vary with the interaction of the two variables and its start/stop times may mislead and suggest unrelated events. Participants were confused by several such cases and it seems CUSUM plots are not very informative in distinguishing overlapping changes.



## CUSUM Diagnosticity: Overlapping Changes

**Figure 9 - Example of a CUSUM chart showing multiple overlapping changes. From this chart it is not clear whether changes in CUSUM slope (December, April, September, June, December) show unrelated changes or fewer common persistent changes. The same data is processed with RE charts in Section 4.4.1.**

The third challenge was related to the ‘coherency’ data validation tactic mentioned in Section 2.5.2 and the multiple purposes of models described in Section 2.5.5. Changes in CUSUM chart slope are ambiguous in that they reflect only system behaviors *whose structure has not been modeled*. Without clear information about which behaviors were represented in the model (e.g. Figure 7), participants’ tendency to explain CUSUM chart changes in terms of simultaneously occurring influences were not consistently correct. For example, they explained CUSUM changes in terms of:

- Cold weather, for a model that accounted for outdoor temperature. Both the energy analyst and specialist at the hospital offered this explanation.
- Weekends and holidays, for a model that accounted for weekends but not holidays. This was complicated by the CUSUM chart date axis not labeling weekends nor holidays.
- Production of a particular type at the factory, for a model that accounted for the reported production quantity. Site A’s energy specialist noted that "when this plant area is shut down, we get savings", and correctly inferred that the model might not correctly account for the energy intensity of that product.

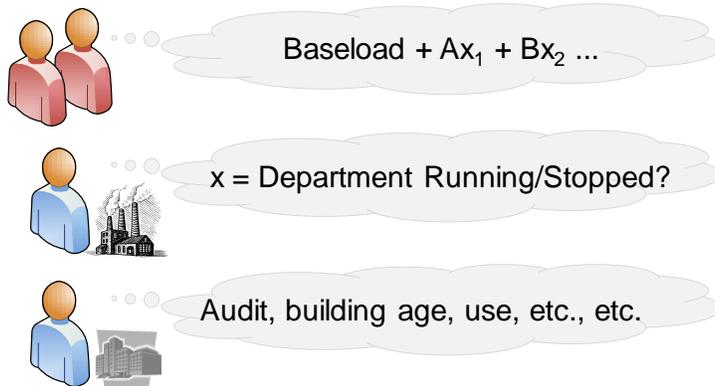
Since the CUSUM (and Control) chart plots only the residual model error (Section 2.3.2), the plots alone cannot distinguish changes in the system or model. Changes to the model observed in the study included driver data missing, mis-measured, or differently processed, and models being deliberately updated by analysts. This ambiguity in chart interpretation was commented on by the hospital energy specialist in an interview, where he remarked:

*“it’s not a matter of being fed little chunks of information, “use chunk X to look at graph Y and it’ll give you answer Z”. The whole thing is a bit more kind of holistic. You have to think, you have to think ‘system’. You know?”*

### 2.5.7. Model Understanding

Since the M&T tool conveyed little information about model development (Section 2.5.5), it is not surprising that energy specialists reported inaccurate understanding of M&T models.

Analysts, by contrast, had no such problem. They had been educated in statistics and developed correct understanding by analyzing data and developing the M&T models. Even if analysts did not know details of a particular model that they may not have developed, in interviews they clearly explained the principles (and limitations) of linear regression modeling.



**Figure 10 - Illustration of different explanations of energy models offered by Energy analysts (top), the Site A factory energy specialist (middle), and Site B hospital energy specialist (bottom).**

Client energy specialists correctly understood that models were based on historic performance. The energy specialist at the factory agreed with analysts on the importance of selecting driver variables based on an engineering understanding of energy-consuming processes. They interpreted baseline energy models as meaning “if everything's operating the way it should... with regular conditions, this is how much [energy] you should use”, and concurred with the hospital energy specialist that a baseline model could be used to reveal the effects of energy projects or changes in production.

However, both client energy specialists admitted to not knowing much about the exact method used to fit energy models. When encouraged to explain their understanding, they presented two different misunderstandings. The factory energy specialist speculated that models were

calculated by an analyst selecting a base load, then specifying "chunks of energy consumption" for each "department" of the manufacturing facility. This is consistent with their explanation of energy use as "normal" with respect to which parts of the factory were operating. The hospital energy specialist thought modeling was similar to industry benchmarking (as reviewed in Section 2.2.6), taking into account not just long term historic data, but also a site audit of major equipment, building age, space utilization, and normalized per square foot of building space.

Both energy specialists described more complicated modeling practices than were or are used for M&T. By contrast, energy analysts described models more conservatively, as:

*"It's our best attempt to try and predict what's occurring on site. It is not always perfect... the model is doing the best it can with the data it has"*

*"[the model is] taking your consumption and then with all the known externalities we have, making our best guess at what the number should have been... to give you something to compare against"*

### 2.5.8. Thought Experiments

Energy specialists' limited understanding of models is concerning because it could reduce how effectively they can apply M&T software for thought experiments. Only one sustained attempt at judging a 'what-if' scenario was observed in the field study. The hospital energy specialist tried to determine from energy data if the recent renovation meant the hospital at risk of overloading its heating system in case of extreme cold weather (Hilliard & Jamieson, 2014b). This reasoning was done without referring to the energy models, even though the gas consumption model had an outdoor temperature parameter and should have been applicable.

### 2.5.9. Social dynamics

The field study took place in hour-long weekly visits, so it was not possible to observe evidence of social dynamics within either site in detail. Some self-reported comments from energy specialists are reported here to illustrate the environment in which M&T tools must function. Some more detail on social dynamics in a wider range of businesses performing M&T is reported elsewhere (Fawkes, 1986; Hilliard et al., 2009; Russell, 2009).

Participants were concerned about interfering with colleagues by demanding their time or attention unnecessarily. At the factory, the energy specialist was reluctant to use the M&T information system in front of colleagues at production meeting, citing a preference for paper printouts that would not risk wasting time with technical glitches. In the hospital, the energy specialist felt "5 minutes is a long time" to interrupt building operation staff, and hoped that analysts would suggest quick specific diagnosis actions such as "I think you have a problem with \_\_\_\_\_, go out in the plant and look for \_\_\_\_\_". The hospital energy specialist had more power to delegate than the factory's, but was concerned about the social constraints of union-mandated scheduling rules. Without day-ahead predictive data based on forecasted conditions (e.g. weather), hospital facilities staff could not be scheduled to start and stop equipment at energy-appropriate times.

Both energy specialists seemed comfortable with the social arrangement of a remote energy analyst advising them and supporting engagement with executive management. However, they seemed to need M&T information system features to help credibly engage peers (at the factory) or efficiently delegate (in the hospital).

## 2.6 Discussion: Practical Gaps in M&T tools

Thus far this chapter has introduced the practice of M&T, reviewed instructional literature on performing M&T (in particular the CUSUM-based model-comparative method) and summarized findings from a field study of M&T practice. The field study revealed naturalistic behaviors rarely or not discussed in the instructional literature. The field study found evidence that existing tools do not adequately support problem-solving, specifically assessing data quality, understanding models, and informing preliminary diagnoses to facilitate social interactions with front-line workers. This section discusses four resultant motivations for the Cognitive Work Analysis of M&T performed in Chapter 3.

### 2.6.1. Software needs for management vs. problem-solving

Management practices of M&T described in Section 2.2.1 require different units and timescales than front-line work (Table 2) because measuring, managing and motivating are different from problem-solving. In the field study the M&T software system was effectively used for record-keeping management tasks. For example, the factory energy specialist used meter data time-

series charts to complete logbooks of equipment start/stop times. At the hospital, M&V models were used to track the financial performance of a major heating system retrofit, although this was being managed by the off-site energy analysts not the on-site participants.

**Table 2 - Comparing context that might be considered by colleagues from different business areas. Such varying perspectives will be categorized in Work Domain Analysis (Section 3.3)**

	<i>Managerial</i>	<i>Financial</i>	<i>Technical</i>	<i>Practical</i>
<b>Concepts</b>	Strategic relevance	Money (\$)	Scientific units (kW, kWh, GJ)	Recalled activities
<b>Timescales</b>	Comparative timescales	Cyclical financial quarters	Ordinal time (minutes, seconds)	Subjective time (shift schedules)

By contrast, the M&T software tool's energy meter data and CUSUM charts were less sufficient in helping participants distinguish, diagnose and problem-solve. For example, since neither site had extensive sub-energy metering, the only way for participants to get more detail about a CUSUM chart trend was to inspect at finer timescales. But even this basic diagnostic action disoriented participants, since the M&T software used timescale-dependent units that changed when users zoomed in to higher frequency meter data (Section 2.5.3). During the field study participants were only effective at identifying known activities, and on-site participants did not seem to have detailed enough understanding (nor time available) to apply a technical approach. If these observations are typical, it is a clear case for improving M&T tools to explicitly support diagnosis and problem-solving work.

## 2.6.2. Assessing data and models

To use evidence in problem-solving requires assessing its trustworthiness. In my preliminary study through participant observation, I spent much more time than I expected assessing and manipulating historic energy data records. Similarly, field study participants were observed to spend about a third of their task time assessing data and models. Some metering software vendors recognize data validation as a problem (Capehart & Capehart, 2005, Chapter 33), but validating data was not well-supported by the observed M&T tool (or the others I had used).

Furthermore, assessing the meaning of M&T models was not described in the literature, aside from a formal approach of learning linear regression modeling. Non-expert participants in this study exhibited signs of having naive mental models (Section 2.5.7), similar to those reported in the residential field studies that inspired this work (Section 2.1.1). The linear regression M&T

models already in use could have been applied for the thought experiment observed in the field study (Hilliard & Jamieson, 2014b), and similarly could be re-used for forecasting. However the M&T tool was not designed to anticipate these applications, and the tool did not expose the model for users to re-use. Lack of understanding of models also limits what management work can be effectively performed. Recordkeeping may be an easy task but if site energy use is highly variable, "convincing people that they have actually saved energy when they have actually used more, and spent more, can be difficult" (Fawkes, 1988, p. 312). Being able to communicate comparisons to a hypothetical "business-as-usual" requires understanding what the comparative model means. This seems a difficult concept to explain, especially using M&T software that has few to no features for summarizing data or model quality. Supporting this crucial part of M&T work, or alternatives to data- and model-dependent approaches are opportunities for tool improvement.

### 2.6.3. Comparison to practice described in literature

In this field study I observed fewer distinct approaches to M&T than described in the literature (e.g. Section 2.2). Participants relied mainly on meter data and CUSUM charts, which is a limitation of the field study. One reason for lack of approaches such as good housekeeping (Section 2.2.3) is that neither site had much employee participation in energy efficiency besides the energy specialists. While participants showed they knew CUSUM interpretation rules described in the literature (Section 2.3.5), participants drew few specific conclusions from inspecting charts. Energy Analysts did not know on-site conditions, and Energy Specialists expressed uncertainty and misunderstandings about models. Both issues are related to structural changes in sites, a feature of M&T work environments under-discussed in the literature (c.f. Cmar & Gnerre, 2005, p. 411). Heavy industrial energy managers reported similar difficulties:

*"Things change in our plant – we'll reroute a pipe, change a meter... the structure of how I calculate the numbers changes. ... I've got to have flexibility. And that's the flexibility I find is not built into a lot of [energy management information] systems" – Participant D (Hilliard et al., 2009)*

Both the factory and hospital experienced non-actionable changes during this field study causing models to fall out of date within months (Section 2.3.3). While waiting for enough data to re-

calibrate models, CUSUM chart perceptual features (Figure 5, Figure 8, and Figure 9) made standard interpretation rules less diagnostic and charts more ambiguous.

Participants' comments about social interactions were consistent with critiques of retrospective data analysis inducing conflict-inducing blame (Hooke et al., 2004). However, I believe inability to develop an initial diagnosis was a larger barrier. Analysts expressed reluctance to bother clients to ask for contextual information to help them diagnose meter data, and energy specialists were reluctant to impose on their colleagues without having an informed guess or a clear course of action. M&T software should support enough different problem-solving approaches that users can work around barriers such as out-dated models or skeptical colleagues.

#### 2.6.4. Understanding and supporting diagnosis work

Participants had difficulty diagnosing changes in energy consumption observed during the study (Section 2.5.6). Besides the large known changes (the factory production schedule change and hospital renovation), participants did not diagnose any energy performance changes during the 8 week study. While some literature grapples with the specifics of diagnosing changes in energy consumption (Bobker, 2004; Lehrer & Vasudev, 2010; Stuart et al., 2007), most literature leaves diagnosis implicit or un-elaborated:

*With an EEMS, [diagnosing] simply involves a straightforward visual analysis of the data, quickly digging deeper into the information where anomalies are evident to uncover problems and gain real insight (Cmar & Gnerre, 2005, p. 417)*

*looking at the same points on the scatter diagram ... should enable someone with technical knowledge of the process to infer the nature of the fault" (Carbon Trust, 2008, p. 15)*

*the person who is responsible for the process [should] be assigned the tasks of collecting energy and process data and explaining the performance in real time.*

*Alternate personnel are not likely to have the required knowledge" (Hooke et al., 2004, p. 38)*

*Understanding the scope for [behavioral] savings requires an awareness of a site's base load and energy usage profile, which can be obtained from advanced metering data. This information can be combined with an understanding of how employees use energy across the business to identify possible savings. (Carbon Trust, 2007, p. 8)*

*Use data visualization techniques to help you and your colleagues find the causes of excess consumption." (Carbon Trust, 2008, p. 19)*

*performance reports can act as a stimulus for investigation and identification of the root causes of both good and poor performance [and] promote operational best practices by eliminating the root causes of poor performance" (Efficiency New Brunswick, 2010, p. 8)*

Information support designers cannot anticipate every possible cause of energy waste, and to some degree "workers [must] finish the design" (Rasmussen & Goodstein, 1987). How much (or little) diagnosis support the information system provides is a design opportunity to reduce the cognitive cost of energy efficiency (Section 1.2).

### 2.6.5. Conclusion

The M&T tools observed in this study supported record-keeping management actions to quantify costs, confirm known events, and in principle would have sufficed to allocate social pressure. For systems with a slow pace of change or very high energy costs, M&V can evaluate capital expenditures or custom control systems can automatically control energy-consuming equipment. However, for investigative operation and maintenance work in systems with a fast pace of

change, existing M&T tools were observed to be inadequately informative. Participants were not effective at diagnosing changes or deriving control actions using a contemporary M&T tool. Participant observation found two commercial tools no more effective at supporting diagnosis.

Designing better M&T tools that remain simple and cheap seems a difficult design challenge. Practical needs that seem un-met are to 1) support diagnosis by 2) aiding users in forming a useful, correct understanding of models and data, and 3) help flexibly delegate or switch to other approaches to problem-solving. As an intermediate step to deriving design opportunities, in the next chapter I will frame literature and field study findings in terms of Cognitive Engineering concepts found relevant to problem-solving in other domains.



## Chapter 3 M&T from a Cognitive Work Analysis perspective

### 3.1 Motivation

This chapter describes the challenges of Monitoring and Targeting (M&T) work in Cognitive Engineering terms through a Cognitive Work Analysis (CWA) of M&T. Three objectives motivated this analysis:

- 1) Exercise the CWA theoretic framework on a novel domain
- 2) Communicate insights from literature and field study (in Chapter 2) to Cognitive Engineering practitioners
- 3) Discover design opportunities and communicate design requirements for cost-effective general-purpose M&T work support tools (for Chapter 4)

As argued above industrial energy management and M&T in particular have not been described from a human factors perspective (Section 1.1 and 2.1). The M&T task is distinctive in that it is:

- A task performed in a wide range of (more or less) engineered domains
- Guided by causal and intentional criteria (Rasmussen et al., 1994); energy consumption is governed by physical laws, but value of energy-related services is judged by workers and managers
- Performed in order to change the (causal) structure of the work system
- In part a diagnosis task

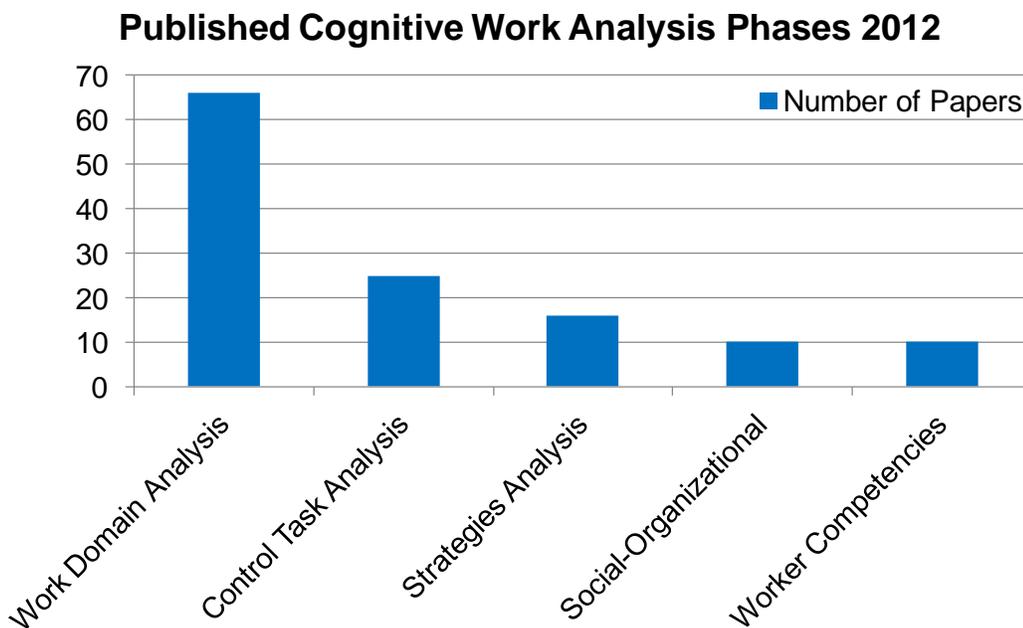
Some of these properties are shared with domains that have guided development of Human Factors theories and methods, for example Nuclear, Aviation, Risk Management, and Quality Control. I compare M&T with these domains in more detail in Appendix A.1.

### 3.2 Methods: Cognitive Work Analysis

First, I very briefly describe the theory and methods of CWA as I applied them to Energy M&T. The summary is structured according to the goals of 3.1 to demonstrate CWA theory in the novel domain of M&T, to describe M&T work (Chapter 2) in cognitive engineering terms and to inspire design opportunities (Chapter 4). This section describes:

- 1) A brief theoretic overview of M&T in terms of control theoretic principles
- 2) Work Domain Analysis (WDA) – “What phenomena are meaningful in M&T”
- 3) Control Task Analysis (ConTA) – “What needs to be done”
- 4) Strategies Analysis (StrA) – “How can work be done”

The analysis develops more detail in ConTA and StrA phases because M&T’s work domain is not well-defined or stable and therefore does not lend itself to WDA. Few published applications of CWA have included task or strategy analysis (Figure 11) (Hassall & Sanderson, 2014, p. 221), and existing methods required some development to tractably apply to M&T.



**Figure 11 - Most published Cognitive Work Analyses have not described Strategies (Hassall & Sanderson, 2014, p. 221)**

### 3.2.1. Energy M&T as Control

To help orient discussion of M&T, a control-theory perspective can be considered. This reflects that M&T is a measurement task and that it can be performed by a variety of human and automated actors. Vicente (1999) suggests considering four canonical conditions for system control (Ashby, 1956):

- 1) The controller must have a goal or goals (e.g., to maintain the set point)
- 2) The controller must be able to affect the state of the system,

- 3) The controller must be (or contain) a model of the system, and
- 4) The controller must be able to ascertain the state of the system.

To illustrate how these conditions apply to M&T in the context of controlling business energy use, and which phenomena they highlight, I present some brief examples.

#### 3.2.1.1. Goals for M&T

The least variant goal of M&T work is presumably ‘cost-effectiveness’. However, management must balance energy efficiency against other strategic business goals (Section 2.2.1). Cost-effectiveness requires quantifying anticipated future savings under varying discount rates, and comparing it against expected labor costs. Some examples of goals or targets in M&T are (Hooke et al., 2004, p. 9):

- budget (whatever we did)
- best practice (whatever they do)
- benchmark (whatever we’ve been able to do before).

We consider goals of energy M&T as overriding purposes in Work Domain Analysis (WDA) and in task-specific goals in Control Task Analysis (ConTA).

#### 3.2.1.2. Affecting the state of energy consumption

Affecting energy use in a system is easy – turn everything off! Affecting energy use without contradicting other goals is more difficult. Depending on the unit of analysis, energy managers can control business energy use at multiple scales from micro to macro (Table 3).

**Table 3 - Three perspectives on how energy can be controlled in organizations, with respect to Work Domain (Section 3.2.2) terms and social activities**

<i>Frame of reference</i>	<b>Physical Functionality</b>	<b>Purpose-related Functionality</b>	<b>Social-Organizational</b>
<i>Examples</i>	Turning equipment off	Scheduling & coordinating equipment use	Reporting
	Maintaining equipment to reduce power draw	Matching energy-intensive service level to need	Educating
	Replace equipment	Modifying business processes	Persuading

Energy consumption controlled by colleagues or customers cannot be micro-managed by lone energy managers with just physical functionality (Table 3). Thus social aspects of control are particularly important. CWA distinguishes relevant system structures in WDA, and social effects are treated separately (although not in detail).

### 3.2.1.3. Models of system energy performance for M&T

Since M&T is a) usually performed on at least a whole-business system (or on particularly large energy-consuming equipment), b) not the only goal of a business, and c) often dependent on social collaboration, several models will be used to guide energy efficiency decisions. Models can be hard to characterize, but will incorporate understanding of the business from financial, technical, and managerial perspectives. Empirical models are used in some M&T approaches (Section 2.2) to quantify goals for metered energy use. These can vary in sophistication (Section 2.3.3) from historic central tendency to correlative models, to first-principles engineering analyses. They are complemented by practitioners' mental models, which will be explored through WDA and the Strategies Analysis (StrA) of Section 3.5.

### 3.2.1.4. Ascertaining System State in M&T

The last condition for control, ability to ascertain the state of the system, is the most relevant to CWA-based design principles (Vicente & Rasmussen, 1992) and may be the most tractable to support through information system design (in Chapter 4). Energy efficiency as a state variable is hard to determine since energy flows are abstract and imperceptible. Also, while instruments can

measure utility supply, it is difficult to tell whether energy was dissipated productively. Some system features relevant to ascertaining system state as a prerequisite to control include:

- Pace of change in the system and corresponding work of keeping state understanding up-to-date
- Sensor (utility meter) accuracy, reliability, aggregation, and frequency – what system state features does each property help infer?
- Coordination of energy consumption with other data records to develop information about state
- Integration of work colleagues’ qualitative experiences (particularly for socially-distributed systems).

Features that define system state are considered in WDA, while actions to determine system state are modeled in ConTA and StrA phases. Some sensitizing M&T concepts I identified using Ashby’s control principles are listed in Appendix A.2.

### 3.2.2. Method: Work Domain Analysis

The first phase of CWA I applied to M&T was Work Domain Analysis (WDA). WDA was originally developed to describe aspects of mental models “of importance for technicians in diagnostic tasks in the control rooms and the workshops of industrial plants” (Rasmussen, 1979, p. 3). Rasmussen derived categories of mental models from terms vocalized in field work (Rasmussen & Jensen, 1973) and scientific models used in engineering.

Conducting a WDA involves developing a set of purpose-driven physical and functional system representations. The representations are ordered from the tangible and physical to the abstract and purpose-driven, and linked by structural means-ends and part-whole relationships. These representations then represent a ‘map’ of possible mental models and can define a comprehensive set of information requirements for design. They are typically represented in three formats: as a full Abstraction-Decomposition Space (ADS), as a subset Abstraction Hierarchy (AH), or as individual relationships between elements (Rasmussen et al., 1994).

The WDA of energy M&T domains was developed using conventional analysis principles described elsewhere (Naikar, Hopcroft, & Moylan, 2005). Because M&T is applied across many systems, and my design objective was general-application task support (Section 1.3.4), I developed a WDA for a category of systems. The WDA is not as specific about a particular

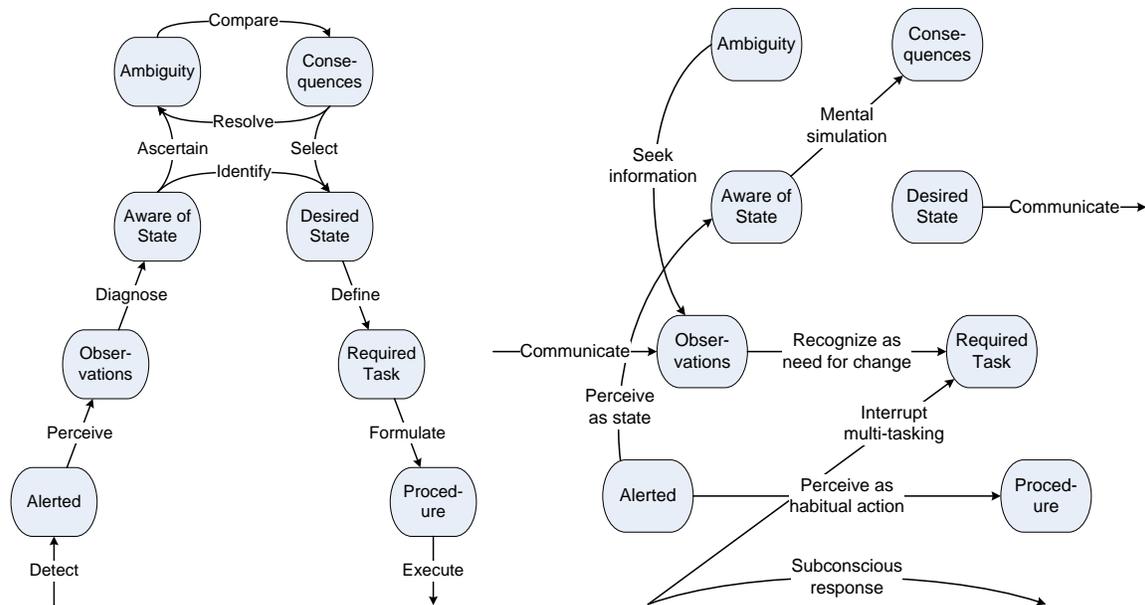
system structure as it could be. For example, rather than listing specific equipment (e.g. valves and pumps in Reising & Sanderson, 2002; Vicente, 1999), it categorizes types of equipment.

Data sources for analysis included literature (Section 2.2 and 2.3), participant observation, and the field study (Section 2.4). The WDA was presented and discussed with energy M&T practitioners but not formally validated. It is presented below in Section 3.3.

### 3.2.3. Method: Control Task Analysis

The next phase of CWA, Control Task Analysis (ConTA) represents work activities in terms of work situations and work functions. Work functions are a template for task sequences, in terms of goals, information products (states of knowledge) and “information processes required to go from one state to another during reasoning” (Rasmussen et al., 1994, p. 65). The usual representation for ConTA, a Decision Ladder (DL), was developed to be “useful for giving a preliminary breakdown of our engineering problem” (Rasmussen, 1974, p. 30), and a taxonomy of increasingly abstract reasoning (Rasmussen, 1974, p. 48). DL notation was intended as a standard framework to communicate psychologically relevant task constraints within systems engineering design teams (Rasmussen et al., 1994, p. 65).

The analysis presented below (Section 3.4) first analyzes activity in terms of stereotypical work situations, or in “work domain terms” (Rasmussen et al., 1994, p. 59), and how these situations influence which activities and functions are active (Naikar, Moylan, & Pearce, 2006). For the subsequent analysis of M&T work functions in DL notation, I drew from canonical CWA texts (Rasmussen et al., 1994) but adopted more contemporary terms for knowledge states (Lintern, 2009), and ‘question-answering’ annotation (Naikar et al., 2006). I adopted this notation (Figure 12) because it is easier to read and annotate (Lintern, 2009) than Rasmussen’s original format (1986) which explicitly separated information encoding, processing, and decoding steps.



**Figure 12 - Examples of Decision Ladder annotation used in Control Task Analysis. At left, normative information-processing steps (lines) between knowledge states (circles). At right, (unrelated) examples of notation for associative leaps and information-processing shunts. Adapted from (Lintern, 2009, p. 65)**

In this analysis, I developed ConTA from three perspectives: as typical activities of Energy Managers (including energy M&T), as an energy control task, and as states of knowledge in diagnosis. The ConTA provided more design guidance than the WDA, and was closely integrated with StrA methods. It is presented below in Section 3.4.

### 3.2.4. Method: Strategy Analysis

Strategies Analysis (StrA) (Rasmussen, 1986) is intended to represent specific constraints on purpose-driven cognitive behavior not captured in WDA (as domain-specific mental model structure) or ConTA (as work functions, goals, and segmented knowledge products). CWA StrA has been less developed in than WDA and ConTA phases, and I found conducting StrA the most difficult phase of analyzing M&T.

StrA modeling aims to capture processes people (or algorithms) use in developing knowledge to pursue goals, including naturalistic behaviors such as flexibly overcoming local difficulties by varying cognitive processes. StrA at least should categorize task-relevant cognitive processes according to distinctions useful for systems design. These distinctions include the perceptual, memory, or experience demands of different strategies (Rasmussen, 1986). Rasmussen

concluded that reasoning processes should be categorized into cognitive strategies so that they "share important characteristics such as

- 1) a particular kind of mental model
- 2) a certain mode of interpretation of the observed evidence
- 3) a coherent set of tactical planning rules" (Rasmussen et al., 1994, p. 70) (numbers added)

There remains a debate over the most useful modeling approach for StrA. Strategies have been described at fine or coarse granularity. They can be modeled as different ways to carry out individual information processing steps between knowledge states in a decision ladder ConTA notation (Kilgore, St-Cyr, & Jamieson, 2008; Vicente, 1999, p. 241) or several aggregated or abstracted task steps (Lintern, 2009, p. 80; Rasmussen et al., 1994, p. 248). Most examples of StrA methods use the former information flow map notation (Kilgore et al., 2008; Rasmussen, 1986), analogous to modeling transfer functions for each individual 'arrow' in DL notation (Vicente, 1999, p. 216). Information flow map notation is minimally defined, and Rasmussen included idiosyncratic symbols to represent for example pattern-matching and system structure. Standardized task analysis notation such as elementary physical operations (Maynard, Stegemerten, & Schwab, 1948) are useful to design manual tasks. A similar method for cognitive tasks, elementary information process notation, has been used in cognitive psychology to model multi-attribute choice decision-making (Payne et al., 1993). Perhaps inspired by this, Vicente (1999, p. 233) suggested adapting such a notation to StrA in the CWA framework.

The approach I followed for this dissertation was a more coarse-grained approach. While fine detail is necessary to validate theories of psychology (Marewski & Schooler, 2011), obtaining it may be intractable for systems design. Instead, strategies can aggregate multiple DL steps (e.g. 'situation assessment' into generalized strategies e.g. 'topographical search') (Lintern, 2009; Rasmussen, 1974), given evidence that they share mental models, observation modes, and tactical rules. Coarse analysis grain has pragmatic benefits of being able to apply stereotypical social strategy templates (e.g. avoidance) to anticipate behavior patterns (Hassall & Sanderson, 2014). This overview of WDA, ConTA, and StrA methods is far from comprehensive but should help interpret the CWA of M&T presented next.

### 3.2.5. Analysis Methodology

The CWA methods above were applied iteratively over the course of participant observation, literature review, and the observational field study described in Chapter 2. The results are one analyst's interpretation of CWA theory applied to synthesize M&T instructional literature and observed / experienced practice in several Canadian settings. Specific limitations of each analysis are described in the results, below.

## 3.3 Results: Work Domain Analysis

WDA is usually conducted on a specific system of interest to analysts. Since M&T is a general-purpose task, intended to be useful in a wide range of energy-consuming business systems, this would require assuming (or selecting) a system to study. Instead, this analysis developed a categorical WDA, similar to examples used to illustrate CWA theory (Rasmussen et al., 1994). To apply this WDA to support work in particular systems, WDA categories could define database structures leaving the specifics of a particular system to be completed (and maintained) by workers.

This WDA is presented according to the three relations used to differentiate mental models of system elements: abstraction, decomposition, and causal/topographical links. Only aspects of the WDA that communicate particularly design-relevant findings from the literature or field study will be discussed. Causal/topographical links were less examined in this analysis since I found them less generalizable aside from types of functional topography (e.g. wires, pipes) and abstract principles (e.g. 2<sup>nd</sup> law of thermodynamics, energy flows from high to low quality).

### 3.3.1. Abstraction and existing M&T knowledge bases

I found the canonical 5-level WDA abstraction taxonomy (Lintern, 2009; Naikar et al., 2005) adequately described types of knowledge used in M&T, with a few exceptions discussed below. First, I summarize the five abstraction perspectives on domains, and then present an Abstraction-Decomposition matrix in Table 4.

The Functional Purposes of a particular business are highly relevant to 'pruning' which system elements should be analyzed. While 'save energy' might seem a purpose of any domain where M&T is performed, I omitted it since a) energy efficiency is rarely a core purpose of an

organization, only a sub-part of financial purposes, b) it serves as a design reminder to relate efficiency to core purposes, and c) it is self-referential. By including only the business's core purposes, 'energy waste' emerges implicitly as functions being used without achieving system purposes, like Rasmussen's 'operator vs. saboteur' analogy (1990, p. 22). Specific Functional Purposes can be found from a business's policy documents.

Abstract functions, values and priorities will include energy. Energy balances are a canonical example of characterizing a system from this perspective (e.g. DURESS II in Vicente & Rasmussen, 1990). Similarly, an energy manager at a heavy industrial site described his data analysis in terms of "reconciling Mass and Money" (Hilliard et al., 2009). Other similarly abstract system representations are in terms of financial analysis, thermodynamics, or other first-principles engineering studies. Abstract representations are laborious to develop, but can show opportunities to reconfigure the system for optimal efficiency (e.g. Pinch Analysis, CanMET Energy, 2003).

Purpose-Related (or Generalized) functions describe the system in terms of processes, for example heating, ventilation and air conditioning (HVAC). These descriptions are somewhat generalizable across businesses, so language can be used for industry benchmarking (Hooke et al., 2004, p. 56). Crucially, comparative analysis strategies (discussed below in Section 3.5.4) often develop models in Purpose-Related functional (PrF) terms, for example with variables that "have physical significance to the actual heat loss/gain mechanisms that govern [business] energy use" (ASHRAE Guideline Project Committee 14P, 2002, p. 24). A typical PrF representation is a process flow diagram, keeping in mind that they usually show only some of the processes required to achieve system purposes.

Physical Functions are a common characterization of equipment, in terms of interchangeable functions (e.g. lighting, ventilation). The interchangeability of this level of representation lends itself to reasoning about straightforward capital investment measures (e.g. exchanging high-efficiency light bulbs or motors). Representations at this level of abstraction include inventories of major energy-consuming equipment (relevant to inventory-based strategies, e.g. Section 3.5.2.2).

Physical Objects are the least abstract representation of system structure. Examples include the form of a building represented as a floor plan, or a thermal image showing “hot spots”. Physical appearance is particularly useful in M&T because efficiency losses are often due to wear and tear, which functional sensors may not indicate. Physical appearance, sound, and condition of equipment can serve as a ‘catch all’ cue source. Representations at this level of abstraction can be used for evaluating service levels (are light or heat needed in this space?) and can be gathered as required (e.g. surveying with a camera). Examples of these levels of abstraction are described in Table 4.

**Table 4 - An Abstraction - Decomposition space for Energy Management in large enterprises. Functional purposes are those of the business, not ‘save energy’. Abstract Functions, Values/Priorities limited to energy-related, and only energy-relevant equipment included.**

	<b>Organization</b>	<b>Section</b>	<b>Subsection</b>	<b>Component</b>
<b>Functional Purposes (FP)</b>	Core purposes of organization. Mission statement. For businesses, typically Profit, Wealth generation, perhaps Social Responsibility	Expanded, breakout description of organization’s core purposes.  Can include energy-consumption-specific parts of core purposes.		
<b>Abstract Functions, Values, Priorities (AF)</b>	Key organizational policies Minimize energy demand to meet purposes Match & coordinate power supply to demand Thermodynamic mass / energy balance summary Net financial balance / flow	Detailed organizational policy First-principles mass/energy analysis of organization, abstract supply, transport transformations, end consumption e.g. “pinch analysis” Economic analysis of organization, financial flows, marginal cost of supply, demand.	Policy fragments (if relevant)  Engineering analysis of each business area’s efficiency & financial payback (e.g. furnace energy balance)	Engineering estimation of individual mass/energy/money balances & flows.  Includes component-level theoretic analyses, e.g. lighting or motor first principles efficiency audit
<b>Purpose-related Functions (PrF)</b>	Aggregated energy-consuming process(es) performed. e.g. “running / idling” Production / working capacity, consumption potential.	Key parts of energy-consuming processes, at more detail. High-level Process Flow Diagram, mostly independent of implementation details	Comprehensive description of energetic processes. Heating, cooling, moving, etc. Sequential relationships between processes.	Detailed parts of energy-consuming processes (if relevant), e.g. evaporative cooling, sensible cooling.
<b>Physical Functions (PFn)</b>	Organization described in terms of total equipment and building functionality For example, real estate portfolio, sum lighting capacity.	Major functional sections, described in terms of purpose-independent functionality, e.g. ability to pump, light, heat, cool. Equipment assemblies, major building areas.	Equipment functional units, described as engineering P&IDs. Building areas, in terms of functional capability e.g. storage annex, 3 <sup>rd</sup> floor office block	Complete inventory of energy-related equipment, in terms of type and capability, e.g. nameplate power, lightbulb types, ventilation fans, pump flow/pressure.
<b>Physical Objects (PO)</b>	Property lots, latitude/longitude.	Map of business floor plan, building wings. GIS Data. Location, condition and appearance of key physical equipment assemblies.	Details of shape, layout of building areas. Detailed physical equipment layout. Appearance of areas, e.g. thermal imaging, video surveillance	Specific condition of individual equipment, asset location tracking. Sights, sounds, smells of space and equipment - Warmth, vibration, leaks, drips, steam etc.

The ADS in Table 4 categorizes information that should be required to reason about business energy performance. M&T practitioners could consider whether such information is available in databases that workers can access, or whether their M&T information system integrates information across abstraction levels (e.g. through unit conversion). Other implications of this WDA are described in Section 3.7. Side-by-side cells in Table 4 are concepts related by abstraction and decomposition, discussed next.

### 3.3.2. Aggregation and Decomposition

The CWA concepts of abstraction and decomposition are particularly relevant to M&T work since diagnosing energy waste requires isolating problems within parts of the system (See Section 2.6.4). The typical five abstraction levels seemed to suit the M&T domain well, but decomposition choices were complicated because of how utility energy consumption is metered in M&T. While business site-level aggregation is always possible (utility bills must be paid), whole-company aggregation is fairly simple (by aggregating sites), and component-level metering can be implemented simply (if at a cost), intermediate decomposition levels depend on specifics of functional structure that will change over time (hopefully as a result of M&T work). A similar difficulty applies to more abstract system structures relevant to reasoning about energy efficiency (Table 5). Sub-meters can be installed on utility lines, but these are usually arranged for pragmatic cost effectiveness, such as short wiring/piping distance. Workers may prefer to reason about energy at different abstraction levels using different aggregations (e.g. Table 4), but these preferences may not match the layout and aggregation of metering that determines what part-whole divisions can be properly accounted for.

As an example of how decompositions vary across abstraction level, electricity consumed in a physical space (e.g. Wing A of a building), is data that conflates various unrelated physical functions (e.g. heating and lighting). Similarly, if a machine's electricity consumption is measured (e.g. a CNC lathe), data may conflate different purpose-related functions the machine can perform (e.g. running, idling, roughing, polishing). More abstract breakdowns of energy consumption might require dynamic mode-dependent aggregation based on some kind of model (at right, Table 5). The variety of models that could be used in M&T reflects the variety of approaches described in Section 2.2 and provides information requirements for M&T software tools.

**Table 5 - How utility energy consumption can be dis-aggregated at different levels of abstraction**

<i>System abstraction level</i>	<i>Ideal decomposition of submetered data</i>	<i>Could be derived using Model of</i>
<i>Functional Purpose</i>	Energy used purposefully vs. wasted	Theoretic energy intensity? Equipment utilization?
<i>Abstract Functions / Values &amp; Priorities</i>	Electric / thermal / kinetic conversion efficiencies? Contribution to quality, timeliness, etc.	Thermodynamic performance of process steps. Energy $\leftrightarrow$ Priority tradeoff models.
<i>Purpose-related Functions</i>	Energy used for heating, cooling, moving, manufacturing, “running/stopped” etc.	Variables directly indicating sub-processes. Time equipment used for each process step, production mode
<i>Physical Functions</i>	Energy used for different <i>types</i> of functionality, e.g. lights, pumps, ventilation, etc. In context of equipment inventory or system <i>capacity</i> .	Equipment inventory (nameplate power data)
<i>Physical Objects</i>	Energy use by location: e.g. basement, motor control center, lighting panel	Business floor plan, equipment locations

A final aspect of the work domain highly important in M&T but difficult to represent was the interaction of time with system properties. Energy is the integral of power over time, and M&T-related processes take place over different timescales (hours to months). This has particular relevance to model-building (Section 2.3.3) since extensive (e.g. mass, energy) and intensive (e.g. temperature) system properties behave differently when aggregated over space or time. Extensive properties aggregate by summing, but intensive properties by averaging. Averaging intensive properties over space or time (e.g. temperatures from daily weather vs. monthly averages) is an idealization that can distort meaning of M&T models and limit them to a particular timescale and aggregation (Table 5). In the task analysis (Section 3.4), I considered how the models of Table 5 might be developed and maintained to accurately represent the work domain at intermediate aggregation levels. Managing modeling scope is essential for M&T information systems to be cost-effective.

### 3.3.3. Causal/Topographical structure

Causal relationships within particular levels of abstraction and decomposition (cells of Table 4) are system-specific and may also change over time. I did not apply them to the design developed in Chapter 4 so they are presented in Appendix A.3. However, I believe that properly representing topographic structure in M&T software is key to supporting the need for data validation observed in the field study of Section 2.5.2.

## 3.4 Results: Control Task Analysis

This Control Task Analysis (ConTA) covered energy M&T activities (Section 2.4) that any M&T information system would need to support. I present three perspectives relevant to M&T tool design:

- Typical situations in which M&T is conducted (Naikar et al., 2006),
- Energy Management activities conducted in parallel with M&T, and
- M&T analyzed in decision-making terms (Rasmussen et al., 1994).

As a supplement to the analysis in decision-making terms, an accompanying Appendix (A.4) annotates the ConTA with diagnosis-related states of knowledge.

### 3.4.1. Work Situations

Work situations (stereotypical configurations of the work domain) are hard to anticipate across domains, but they're very relevant to energy M&T, particularly as context for developing energy models (Section 2.3.3) or interpreting meter data (Section 2.2.4). Some M&T-relevant examples are:

- Workday / Production
- Weekend / Idling / 'hot' shutdown
- Holiday / Maintenance Shutdown
- Summer, air conditioning season
- Winter, heating season

Such business situations have three main implications. First, they affect which work domain purposes are active and therefore the values that should be considered when controlling

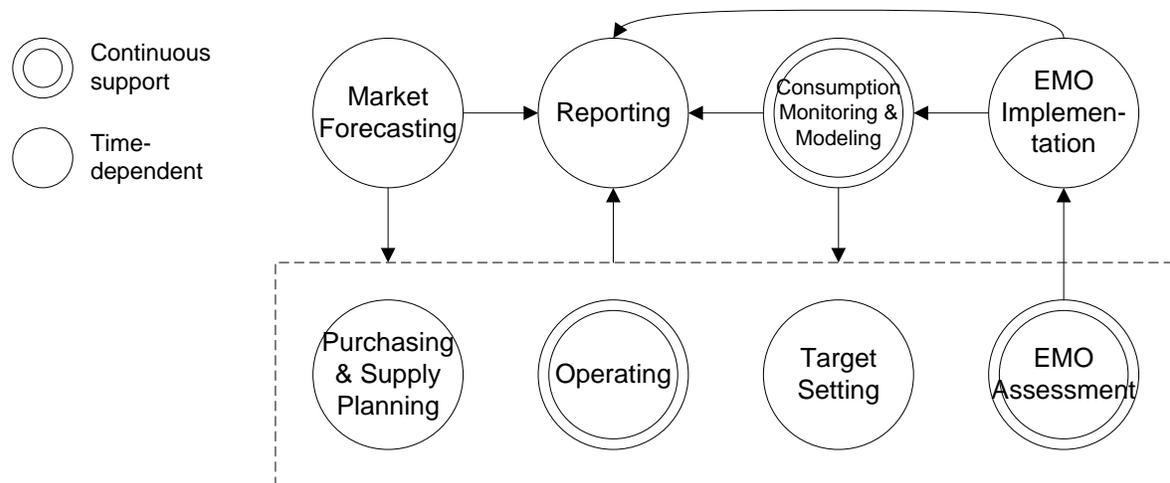
functions. For example during shutdowns, energy control work can focus on idling equipment and leaks, while when operating, efficiency must be balanced with production quality or customer satisfaction. Second, changes between modes / situations are relevant to diagnosing problems by observing whether they persist (e.g. in the Profile Patterns strategy, Section 3.5.1). Finally, business modes are an important consideration in selecting a statistical model training dataset (e.g. for Comparative Analysis strategy, Section 3.5.4), as models will implicitly reflect the proportion of modes they were trained on. Conversely, knowing the modes over which a statistical energy model was trained is necessary to judge what the model represents. Mode-relevant information was identified as a requirement for the model summary tool developed in Chapter 4. This overview is brief, since situations are site-specific. The next ConTA modeling construct, work activities, are more consistent in M&T.

### 3.4.2. Work Activities

Energy management work can be characterized as work activities (which include M&T, as in Figure 1). Even though this dissertation and the work support tools of Chapter 4 focus on M&T, related work activities describe the context in which the M&T function will be performed.

I developed a set of Energy Management work activities from literature review and qualitative interpretation of the preliminary interview study (Section 2.1.2.1), shown in Figure 13.

Comprehensive Energy Management programs include most of these activities (BRESCU, 2001). The M&T energy control work addressed in this dissertation comprises mostly Consumption Monitoring & Modeling, with some elements of Target Setting and Reporting.



**Figure 13 - Some typical functional activities of Energy Management (Hilliard et al., 2009). Energy M&T represented mostly by "Consumption monitoring & modeling" with some "Target setting". EMO stands for Energy Management Opportunity. Arrows represent information flow between activities (Table 6)**

The importance of M&T to other energy management activities is illustrated in Figure 13.

Developing an understanding of energy consumption through Consumption Monitoring informs purchase planning, operating, and continuous improvement (e.g. ISO Technical Committee 242, 2011). Information flows between work activities are shown as arrows in Figure 13, and examples of communication between activities are outlined in Table 6.

**Table 6 - Example information flows between energy management activities (Hilliard et al., 2009). Work activities of Figure 13 are shown as topmost diagonal cells and information flows (as questions) in the tabular intersections between functions.**

<b>Market Forecasting</b>							
Where are energy prices going?	<b>Reporting</b>						
-	What / where are we consuming?	<b>Consumption Monitoring &amp; Modeling</b>					
-	Success stories?	How much was saved?	<b>EMO Implementation</b>				
Projected prices over life cycle?	Payback, ROI, NPV? Disruption, Risks?	What areas have greatest potential?	What are the work orders?	<b>EMO Assessment</b>			
How much must we improve?	What are tough / achievable targets?	What is historical benchmark?	-	How great are our potential savings?	<b>Target Setting</b>		
Anticipated spot prices? Demand Response?	Production? Quality? Costs?	Over-spend areas? CUSUM meaning?	-	How much will this complicate our job?	-	<b>Operating</b>	
Trends in spot prices?	How did you save us money?	Correctly billed? Typical qty. / peak?	-	-	-	-	<b>Purchasing &amp; Supply Planning</b>

The Consumption Monitoring activity (3<sup>rd</sup> column in Table 6) informs situation assessment in M&T and relates to most other energy management activities. It is informed by and contributes to other work activities such as:

- Reporting utility consumption for business management processes (e.g. accounting), ideally disaggregated by technical or management criteria (Capehart, Turner, & Kennedy, 2008, p. 18; Technological Economics Research Unit, 1979).
- Implementing Energy Management Opportunities (EMO), by measuring and validating (M&V) anticipated energy savings (ASHRAE Guideline Project Committee 14P, 2002)
- Assessing potential EMOs, by identifying functional / physical areas of high or increasing consumption

- Target setting policy deliberations with executive management, through substantiating what energy performance is feasible
- Operating the business, by identifying dis-aggregated areas where energy consumption exceeds targets (over-spend), and supporting operation choices (e.g. work scheduling).
- Energy purchasing and supply planning: verifying billing quantities, forecasting utility consumption and maximum power/rates

These links illustrate that good M&T is important because it develops data and understanding that many other activities depend on. This also suggests data input and output requirements for M&T information support tools to support a range of approaches to M&T work. The next step of ConTA was to analyze the ‘consumption modeling and monitoring’ work function in more detail using Decision Ladder notation.

### 3.4.3. Decision Ladders of Energy control and Data cultivation

Since I observed data record-keeping and model-assessing competing for time with M&T behavior (Section 2.6.2) I chose to analyze monitoring and modeling as two separate (but closely related) work functions with distinct goals: 1) Controlling Energy Costs and 2) Cultivating Data & Models. The two functions are shown as decision ladders in Figure 14, and the states of knowledge described in more detail in Appendix A.1. Including both control and representation-maintenance functions in a DL notation is unconventional and has theoretic implications discussed later.

The “Control Energy Costs” work function is intended to represent how a business’s energy-consuming activities can be controlled. Key observations include power draw, utility energy consumption, and cost data. This function can describe workers monitoring shift-by-shift behaviors, engineers assessing structural improvements (EMOs), and in principle how management assesses financial performance. Controlling Energy Costs has the ultimate goal of minimizing net delivered energy cost to meet business requirements, through acting on the natural system.

The “Cultivate Data & Models” work function describes how automated meter systems, log keepers, and analysts collect and document energy performance-relevant observations and models of the system. The term “cultivate” is intended to imply the sustained work involved in

recording data, developing models, and maintaining their coherence as the work system changes. Cultivating Data & Models has the ultimate goal of maintaining useful, trustworthy representations of business energy performance, through acting on instrumentation and data records.

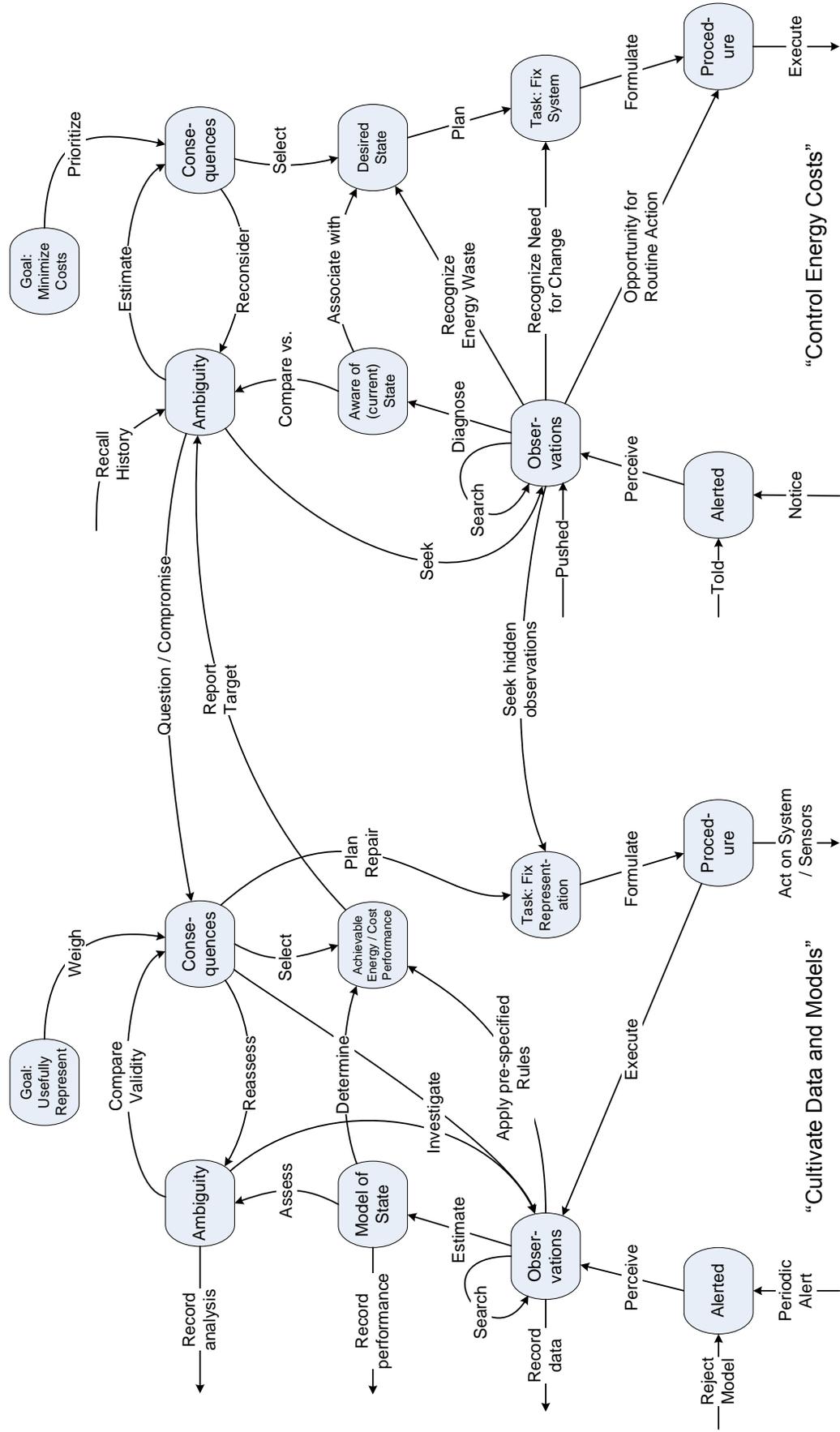


Figure 14 - Linked decision ladders for two M&T work functions: Cultivating Data & Models (left), and Controlling Energy Costs (right). Ovals represent states of knowledge, described in Appendix A.1. Arrows represent information processing steps.

Showing the “Cultivate Data” function as a distinct DL emphasizes that it is only useful as far as it supports the “Control Energy Costs” function through reporting a target achievable energy performance or resolving ambiguity (Figure 14, top). Cultivated data and models that are not applied are unprofitable data hoarding. The need for pragmatism is reflected in the goal of producing useful ‘actionable’ records, rather than an exhaustively complete accounting. Workers performing this function (and information support tool designers) must balance the cost of accumulating ‘data baggage’ versus the benefits to present and future decision-making.

One caveat of Figure 14, discussed in Appendix A.1, is that the left side of both DLs (Alerted, Observations) aren’t clearly distinct between functions. Similar observations of the work system may be relevant to both goals. The ‘Cultivate Data’ function may emphasize quantitative, model-applicable data, while qualitative, ephemeral observations are more useful to ‘Control Energy Costs’. The work functions are characterized in more detail in Appendix A.6 with four examples of how existing M&T practice can be mapped onto the two work functions. Strategies for M&T (discussed below) can be described in terms of how much they depend on each work function.

### 3.5 Results: Strategies Analysis

As the field study suggested difficulties in learning correct mental models (Section 2.6.2) and the importance of diagnosis (Section 2.6.4), understanding strategies for M&T seemed particularly important to designing M&T tools. I carried out the StrA in parallel with ConTA. From field studies (Hilliard & Jamieson, 2014b) and Chapter 2’s literature review (particularly CIPEC, 2010) I categorized<sup>5</sup> six M&T diagnosis strategies. To focus on diagnosis, these strategies cover only situation assessment, the cognitive products and processes shown on the left side of the decision ladder (in Figure 14). Strategies for action selection and execution may be less promising for general-purpose information tool support, since they may be easier-to-delegate maintenance work or more variable engineering design work. Analysis of action selection M&T strategies is an opportunity for future work.

---

<sup>5</sup> I tried coding behavior fragments (Sanderson & Fisher, 1994) to do a fine-grained strategy analysis, however did not achieve inter-rater reliability (Krippendorff, 2004).

### 3.5.1. Overview of M&T Strategies

The strategies that I identified for interpreting business energy performance are listed in Table 7. They can be distinguished by the degree to which they depend on cultivated models and historic data, or whether they can be performed with mental models alone.

**Table 7 - Six M&T situation assessment strategies, categorized by whether they need external data and model maintenance (left) or can be recognition-based (right).**

<i>Situation Assessment Strategies for Control Tasks:</i>	
<i>Use Cultivated Data record or Models to Control Energy Costs</i>	<i>Recognize Energy Cost Control opportunities</i>
Energy consumption and cost analysis	Event – action time-series association
Reconciling equipment inventory with consumption	Consumption time-series profile pattern-detection
Comparative analysis against normative/historic model	Condition survey or ‘Good Housekeeping’

Differences between strategies are important for tool design since people prefer to switch between strategies to minimize effort and maximize anticipated accuracy (Gigerenzer, 2001; Payne et al., 1993; Rasmussen, 1986). Strategies can be compared by tradeoff criteria (Section 3.2.4), for example as in Table 8.

**Table 8 - Six M&T situation assessment strategies, compared in terms of properties that may be relevant to strategy switching trade-offs**

	<b>Comparative Analysis</b>	<b>Equipment Inventory</b>	<b>Qty. &amp; Cost Analysis</b>	<b>Condition survey</b>	<b>Profile Patterns</b>	<b>Event Association</b>
<i>Frequency</i>	Daily +	Weekly +	Yearly +	Instant	Instant +	Any
<i>Aggregation</i>	By-meters & data	By-meters	Whole-site	By-item	By-meters	By-meters
<i>Statistical Model Complexity</i>	++	++	+	-	+	-
<i>Quantitative Referent</i>	Model (Historic)?	Equipment list	Other business costs	None	Historic data	None
<i>Consequence of Domain Change</i>	Optional Model Maintenance	Model Maintenance	Implicit in next update	None	Wait for 'new normal'	None

These six strategies describe current practice (Section 2.2) and are not exhaustive. Other algorithms (cognitive processes) such as neural networks, expert systems or fuzzy logic approaches have been demonstrated for fault detection and diagnosis (Capehart & Capehart, 2005, p. 290). The comparative analysis strategy in particular can use any suitable algorithm for developing an energy performance model, though tradeoffs of modeling algorithms are discussed in Section 3.7.3 and 3.7.5 below. I will briefly overview each strategy in terms of theoretic distinguishing criteria: how data are interpreted, what type of (mental) models are needed, and (briefly) tactical rules for planning, search, decisions, and stopping.

### 3.5.2. Strategy summaries

This subsection describes my formulation of the strategies listed in Table 7, the first four briefly, then in more detail the remaining two: Condition Survey and Comparative Analysis (the basis of the design intervention in Chapter 4). Tactical rules were not modeled as detailed information flow maps, as I did not have the data to validate process descriptions.

### 3.5.2.1. Energy Consumption and Cost analysis

This strategy is to interpret observed data as the ‘big picture’, at yearly timescales. It requires record-keeping of historic site utility bills. Annual bills are summarized in terms of utility consumption (power and energy) and associated line-item costs (peak demand and bulk consumption). Productive output of the business is interpreted in terms of aggregated measures such as building floor space (physical objects), or manufacturing production. These observations are processed with simple ‘black box’ operations just sufficient to do unit conversion or normalize energy consumption (CIPEC, 2010). Outputs could include bill costs due to energy versus peak power, and simple normalizations of energy in terms of time (e.g. kWh/day, Load Factor, Load Factor/Utilization Factor) or purposes (kWh/ton). The resulting externalized models are sufficient to direct colleagues’ attention to high-cost business areas, compare against other business-relevant costs, and sensitize people to important phenomena to attend to in the future.

The strategy’s purpose is not to inform specific tasks, but rather to educate colleagues with simple models of the energy-related costs of business activities. The intent is for workers to apply their new understanding to infer the consumption and cost implications of equipment and activities that they control. Because this strategy depends on averaging out long-term invariant relationships between energy and production, resulting models may not be very meaningful at short timescales. There’s also a potential risk of discounting base loads or conflating mode prevalence (task situations in Section 3.4.1) with energy performance, especially if using KPI-type production-specific units (Harris 1989).

### 3.5.2.2. Reconciling Equipment Inventory with Consumption

With this strategy (Table 7), observations are interpreted as:

- Energy consumption data at coarse (weekly to yearly) timescale and some particular aggregation (whole-site to individual submeters)
- Corresponding sets of metered equipment in terms of type (physical function), power draw, and run time (e.g. twelve lights, 100W each, active 8 hours per day)

This strategy typically develops a tabular model of the business as an aggregated collection of equipment types, characterized in terms of maximum power draw and percentage run time (load factor). Tactical rules are to iterate between:

- 1) searching out equipment (largest first, until impractical),
- 2) characterizing equipment power consumption and operation time,
- 3) compiling consumption data,
- 4) comparing equipment list with consumption, and
- 5) reconciling mismatches through refining the model (list) of equipment or the compiled energy consumption data.

The strategy halts once the equipment inventory is congruent with actual energy consumption. The Equipment Inventory strategy primarily cultivates data and models. It develops representations that must be maintained as the business changes. Equipment inventories are useful for business management decisions (such as ‘which equipment types account for the most consumption?’ or ‘which are over-sized and operate the least often each day?’). However, when the business system changes, it may be more difficult to draw conclusions that span changed equipment inventories.

### 3.5.2.3. Consumption time-series profile pattern-detection

For this strategy, meter data must be observed at least at an hourly frequency, and often at whole-business aggregation. Consumption data is interpreted as a time series aligned by day-of-week, production shifts, or other temporal reference frames (Nyssen & Javaux, 1996). Data is interpreted, un-processed, as repeating patterns representing site activities or major equipment operation. This strategy does not require formal models of normative or descriptive system behavior. Instead, time-series patterns are used as their own historic referents, to judge ‘normal’ or ‘abnormal’ patterns. Calendars are used as proxies for system activities (daytime vs. overnight, weekdays vs. weekend consumption patterns). Mental models of recalled business activities can be reconciled with energy data. One source describes this strategy as:

*"Look at your graph of energy consumption and think about how this fits in with the pattern of production/occupancy of your building. Investigate any suspicious areas, for example, has the energy use continued at a high rate during periods of low production? Or, is energy still being used during office holidays?"*

*Also, consider other factors which might affect your energy consumption: did you change your production patterns, or did you need extra heating due to cold weather?" (Carbon Trust, 2006, p. 4)*

Tactical rules are outlined in instructional literature (CIPEC, 2010, p. 57), including suggested switches to or from the Equipment Inventory strategy. They note:

*"Interpreting a demand profile is not just science (technical skill) – art (interpretative skill) is involved too. Good knowledge of the facility, its loads, operational patterns and the examples in this section should provide a solid basis for developing that interpretative skill." (CIPEC, 2010, p. 57)*

A time-series consumption analysis strategy can take advantage of workers' detailed understanding of daily work, or un-recorded phenomena, but is not very useful for managing by exception since it requires inspecting day-by-day. It is also less sensitive to gradual increases in power draw typical of badly-maintained equipment. Other strategies can be coordinated to reduce the need for 'artistic' interpretation.

#### 3.5.2.4. Event-action times-series association

This strategy was not clearly distinct, but seemed to be a fallback heuristic invoked from other strategies. When participants appeared unable to reconcile energy consumption with a referent (e.g. equipment inventory, expected schedule) participants sometimes explained it by recalling or speculating some notable property of the time period. This seems similar to the 'take-the-first' availability heuristic (Gigerenzer, 2001). In the field study (Section 2.5.6), participants seemed to fall back on event-based explanations when examining both time-series energy profile patterns and Comparative Analysis CUSUM charts. Even though the meaning of changes in these charts is very different, participants did not seem to distinguish between them.

For event-association strategies to be effectively applied, workers must have a good understanding of which phenomena other strategies can and cannot account for (discussed below in Section 3.5.4).

#### 3.5.3. Condition survey Strategy

Because this strategy describes good housekeeping work highly regarded in the literature (Section 2.2.3) I discuss it in slightly more detail. Unfortunately this behavior was (to my surprise) not observed in the field study. It serves as a contrast to the Comparative Analysis strategy (the ongoing focus of Chapter 4 and Chapter 5)

For the Condition Survey strategy, data is gathered in-person (Fawkes, 1986), using human senses to observe proximal cues such as the sound of compressed air leaks, or observation of equipment running (Kirlik, 2006). Human senses can be augmented, e.g. with infrared cameras or ultrasonic leak detectors. Observations are interpreted as signs directly indicating energy waste or equipment degradation. The Condition Survey strategy does not require externalized models or data records, only mental models of what ‘good’ condition and operation look like<sup>6</sup>. However, diagnosis hunches reached by data-driven strategies can initiate a Condition Survey search.

Strategic rules include skill-based regulation of energy-consuming processes (e.g. unconscious sensorimotor recognition and action such as turning off light switches reflexively), rule-based lists of ‘energy saving tips’ or recalled wasteful situations, and knowledge-based estimation of energy use and associated costs on-the-fly. Example states of knowledge for the Control Energy Costs task (Figure 14) specific to this strategy are listed in Table 9.

---

<sup>6</sup> Because of human tendencies to acclimatize to their everyday surroundings, external auditors or a ‘neat freak’ may be more competent at effectively surveying equipment condition.

**Table 9 –States of knowledge specific to Condition Survey strategy. This strategy can be conducted with mental models alone, if the actor can recognize stereotypical signs of energy waste (e.g. hot air drafts) or site-specific condition or operation problems**

DL Knowledge Category	<i>Condition survey</i> Strategy for Control Energy Costs work function:
<b>Alerted</b> to potential issue	Does that sound / look / feel like energy waste (hiss, light, warmth?) Does that seem like something that consumes a lot of energy? Does a colleague want me to investigate something? Is there yet more equipment to check?
<b>Observations</b> – Aware of the dimensions of the situation	What is that equipment? What type of equipment is it? What function does that equipment perform? What is the condition of that equipment? Is it old? Worn? Clogged? Are those leaks? How much air / steam / fluid is leaking? How hot/cold is this area / equipment / leak (skin touch or thermal scope) Is the equipment running? Is the equipment doing something productive? What is the equipment's power draw? (nameplate / measured) How is the equipment adjusted/configured? Has all equipment been checked yet?
Aware / Model of current <b>System State</b>	How might the equipment's condition affect its power draw / energy use? How does the equipment's operation affect its energy use? What purpose-relevant services does that equipment function provide? (e.g. is function productive?) How might the equipment's power draw and operating hours combine to total energy use? How much energy / stuff is being wasted? What are the costs of energy waste? Equipment repairs / actions?
<b>Ambiguity</b> – Aware of potential system states	How typical is the particular system state observed now? What observation errors could have been made? Is it economic to correct equipment condition? / operation? Is it economic to reduce demand for services?
Aware of <b>Consequences</b> of potential states	Will changing equipment operation / condition reduce waste? How could omissions / mistakes in taking observations be misleading? What might go wrong if this equipment operation is changed? What might go wrong if this equipment is replaced?
Aware of <b>Desired / Achievable State</b>	What condition should equipment be in? What equipment should be repaired? What equipment should be replaced? When should equipment be running?
Aware of <b>Required Task</b>	What changes are involved in repairing leaks or equipment? What changes in work practices need to be made to match operation to demand?
Aware of <b>Procedure</b>	What sequence of actions need to be carried out to repair the leak / equipment? What steps need to be followed to operate equipment only when needed?

Some characteristics of the Condition survey strategy are:

- It relies on difficult-to-quantify cues that are easier to sense in-person. Workers may be slow to switch to this strategy if they are monitoring energy from an office desk.
- When on walkabout, suspicions about energy waste may not be actionable and may require note-taking and later investigation by switching to a different strategy.
- Observations can be directly associated with a diagnosis and action, but unlike data-driven strategies do not quantify energy/cost waste.
- Condition survey practice is probably subject to sensory bias, as observed in folk energy management (Kempton & Montgomery, 1982). Novices may overestimate energy spent in tangible / proximal forms (light), underestimating the intangible / distal (heat). This is an opportunity to augment human senses (e.g. with thermal infrared cameras).

The Comparative Analysis strategy is discussed next to highlight how it may particularly compliment Condition Survey strategies.

### 3.5.4. Comparative Analysis Strategy

This strategy was the most well-discussed in the literature (Section 2.3) and the most frequently observed in the field study (Section 2.6.3). Comparative Analysis combines the time-series pattern-recognition of the consumption profile strategy (Section 3.5.2.3), with the benefits of normalizing energy consumption to account for uncontrollable disturbances or productive energy use (Section 3.3). Because of its popularity and its use of statistical models, it forms the basis of the M&T work support tool described in Chapter 4 and evaluated in Chapter 5.

Comparative Analysis strategies can be done with pen-and-paper, but are usually partially automated (shown as M&T task steps in Section A.6.1). Algorithms interpret energy meter and quantitative contextual data symbolically as inputs to an energy performance statistical model. The residual between model-derived and actual energy consumption is then visualized (such as in CUSUM charts) and interpreted against a mental model as “energy under/overconsumption”, using rule-based recognition behaviors (Section 2.3.5) or if required knowledge-based reasoning (Section 2.5.6). This strategy can use a wide range of computational models of varying complexity (see Section 2.3.3). Statistical system models can include:

- An all-time historical average

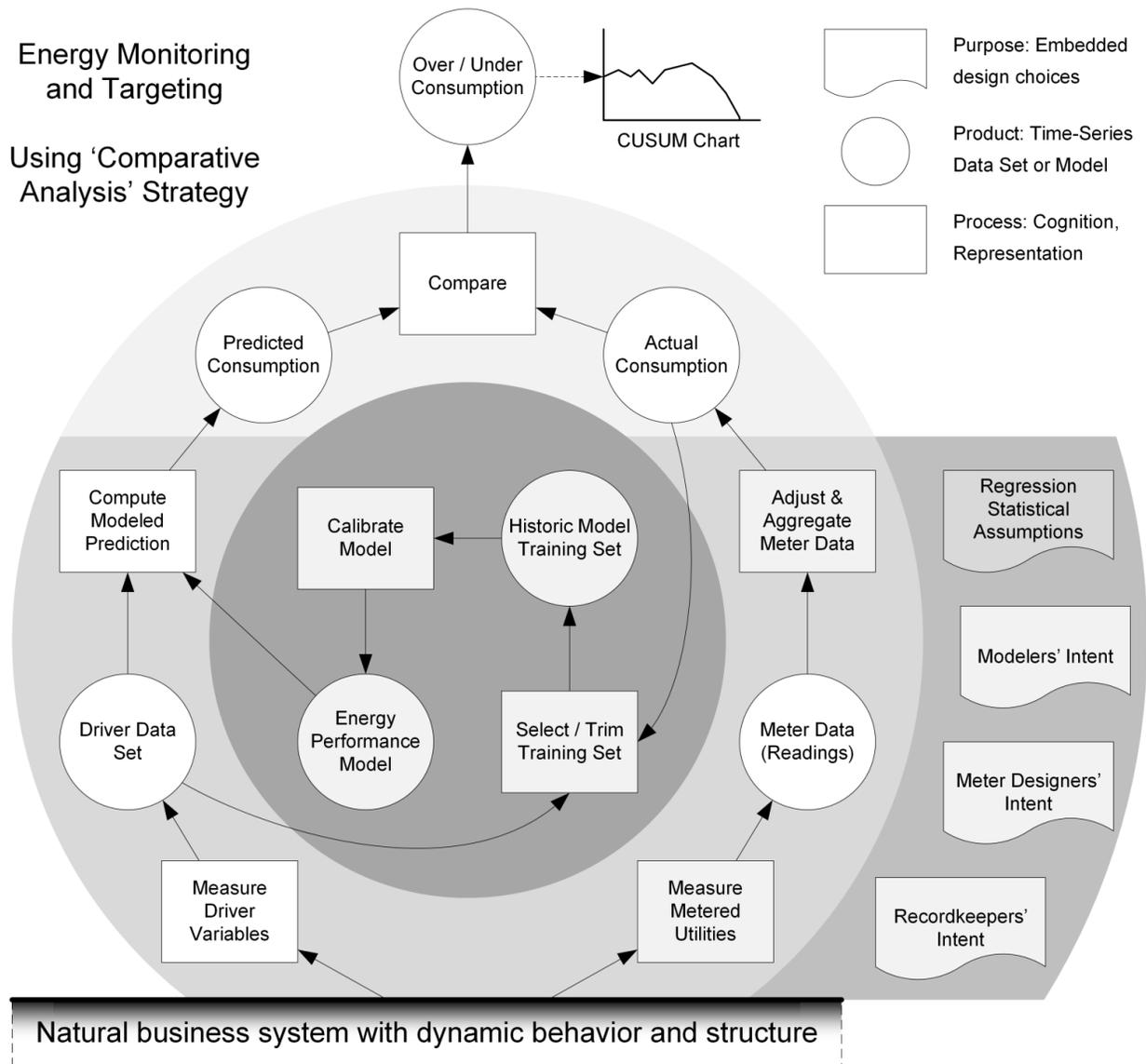
- Piecewise averages assembled from operation in different modes, e.g. weekdays/weekends (Stuart et al., 2007).
- Linear regression on recorded measures of controlled disturbances or productive outcomes (weather, production, scrap, hours of operation) (Harris, 1989)
- Nonlinear regression, or regression on transformed data (e.g. Heating Degree Days) (Fawkes, 1988)
- Contemporary statistical modeling methods (Capehart & Capehart, 2005)

Any of these methods can be used with a Comparative Analysis strategy to transform observed data into information about ‘over/underconsumption’ relative to the statistical model. Simpler statistical models require less data and effort to build and maintain, but compensate for fewer influences on energy consumption. Changing the complexity of the statistical model necessarily affects how complex a mental model is needed to interpret residuals, but the relationship is not necessarily proportional (a design tradeoff discussed in Section 3.7.5). Because Comparative Analysis depends on inferring present (performance) relative to past (model training), it requires both work functions described in Section 3.4.3 and Figure 14. States of knowledge for the Comparative Analysis Strategy are listed in two columns (one for each work function) in Table 10.

**Table 10 - Specific states of knowledge for Comparative Analysis strategy. Unlike the Condition survey strategy in Table 9, Comparative Analysis requires cultivated data and models to inform energy cost control.**

DL Knowledge Category	<i>Comparative Analysis</i> Strategy for two goal-driven work functions:	
	Cultivate Data / Model	Control Energy Costs
<b>Alerted</b> to potential issue	Are data or models unexpectedly suspicious? Are there new energy-related (e.g. production) data available? Utility meter data?	Is there a decision (e.g. Operation, purchasing) that I need to make? Did a surprising energy waste event happen?
<b>Observations</b> – Aware of the dimensions of the situation	What are the meter readings for utility consumption? Electricity? Gas? What energy-relevant data is available? Production? Outdoor temperature?	← Cultivating Model Observations, plus e.g. How much area/demand/production was un-recorded over this time? What activities / events happened at this time? (startups/shutdowns/maintenance)
Aware / Model of current <b>System State</b>	What true energy consumption does the meter data indicate? What underlying energy-related processes do other data indicate? What model most usefully predicts energy consumption based on energy-relevant data?	← Cultivating Model State Awareness, plus e.g. What true productive services were performed with this energy consumption? How normal / typical were the un-quantified phenomena that happened?
<b>Ambiguity</b> – Aware of potential system states	Do energy / -related data still reflect the actual system? Have sensors or records changed? How ambiguous are the energy performance model's calculations? What could residual energy model errors mean? How well-satisfied are the statistical assumptions of the energy performance model?	How does actual energy consumption compare to historic energy consumption? Or Model-calculated energy consumption? Does comparison suggest structural changes in energy performance happened, or persisted? Do un-quantified phenomena adequately explain over/under-performance? Is this actual energy consumption or model-calculated consumption suspicious?
Aware of <b>Consequences</b> of potential states	What are the implications of violating energy performance model statistical assumptions? Should the model be changed to be more useful? Will correcting the energy or energy-related data record make them more representative?	Will it be economic to meet this desired/target energy consumption? Are the costs significant if this energy overconsumption persists? What if observed data is wrong? What if explanatory observations are missing?
Aware of <b>Desired / Achievable State</b>	What achievable energy consumption does a historic / target energy model calculate?	What should the energy consumption have been for the productive output, according to model, history, or arbitrary target? What work activities/conditions are desired to achieve target energy performance?

Tactical rules for the Comparative Analysis strategy can be broken into two steps: 1) the symbolic data processing of a performance metric, and 2) the interpretation of the performance metric. A set of symbolic data processing steps are diagrammed as an information flow map in Figure 15.



**Figure 15 - Information Flow Map for Comparative Analysis strategy, transforming observations (bottom) to a system state indicator (top). Energy models are maintained and updated from time to time (center). Darker shading indicates elements that were poorly documented and less observable to field study participants.**

The normative information flow map in Figure 15 shows the cognitive processes and intermediate knowledge products in developing analysis-derived and actual energy consumption. At left, observations of the business system (driver data) are processed into a model-predicted

energy consumption. The link between driver data (observations) and predicted energy consumption (desired system state) is associative according to pre-computed energy model trained in the steps of the inner circle of Figure 15. The shading is slightly darker grey to indicate that these processes and intermediate products were less observable in the information system observed during my participant observation and the field study of Section 2.4.

The right arc in Figure 15 describes utility meter observations being processed into an ‘actual’ business energy consumption. This requires a model of the functional structure of utility (sub-) meters. When considering whole businesses this can be a simple 1:1 mapping. However if applied to sub-aggregations meters may need to be summed or differenced (see Section 3.3.2 and Appendix A.3), and for gas or fluid meters whose calibration is never perfect, meter data may require scaling adjustments. The resulting actual energy consumption is used for comparative analysis, and (from time to time) for calibration of models on historic empirical data. The output of the normative information flow map of Figure 15 is an energy performance indicator (shown at top as ‘CUSUM chart’). This product is interpreted with subsequent cognitive processes, which could be described by extending the information flow map.

I did not explicitly model interpretive processes, but in field study observations workers rarely took the chart at face value and usually considered ambiguity in the meaning of the computed energy performance indicator (Section 2.6.2). However ambiguity is not easily diagnosed from the performance indicator alone, since it is influenced by every process the whole way up Figure 15. Errors in data collection, violation of model assumptions, and non-ideal model characteristics will all affect the meaning of the performance indicator. Some of the inferences that must be made to appropriately rely on the performance indicator are common to all data-driven strategies and are discussed in Appendix A.5. Others are strategy-specific and require answering questions in Table 10. For pragmatic time-constrained work in business environments, or with un-integrated software tools people may not be able to validate these inferences. From the field study it seemed that people could substitute an understanding of designers’ intent, shown at right in Figure 15.

The Comparative Analysis strategy can support switches to or from with other strategies (Appendix A.7). The CUSUM charts typically produced from Comparative Analysis integrate well with consumption and cost analyses (Section 3.5.2.1), since they quantify accumulated costs

even on slow timescales. CUSUM charts also work well with equipment inventory strategies (Section 3.5.2.2), as they can quantify savings rates (such as M&V of known changes). Undiagnosed overconsumption rates can be compared against equipment that might have been left running. CUSUM charts share typical time-series representations with time-series analysis (Section 3.5.2.3) and event-action (Section 3.5.2.4) approaches. However, CUSUM charts inform Condition Survey (Section 3.5.3) or other topographic search strategies (Rasmussen, 1986) only through isolating energy consumption in time and possibly within sub-metering disaggregation (Table 5). If submetered areas are not economically searchable, more information must be sought, possibly from work colleagues responsible for energy-consuming equipment.

### 3.6 Socio-organizational Analysis

Socio-organizational analysis (SOA) is helpful to consider how task information flows described in Section 3.4 and mental models of Strategies in Section 3.5 can be distributed among workers. Because field study data was limited (Sections 2.1.2 and 2.4), and I did not have a particular social arrangement in mind for the design application of Chapter 4, I briefly discuss only three social arrangements: Centralized, consultant, and worker-engaged. In large organizations at least, interviewees reported a problematic tendency for the M&T task (Figure 14) to be centralized in a lone “energy guru” (Hilliard et al., 2009). Centralized M&T seemed to work well for capital investment, but with risks including:

- Brittle, impermanent energy efficiency programs. If a ‘guru’ leaves the business, the energy management program might end.
- Bespoke M&T information systems designed for one user’s personal use. If the “guru” left, data and models might be un-maintainable.
- Dependence on colleagues for information about un-instrumented energy-wasting conditions or behaviors that were only known locally.
- Dependence on ‘employee engagement’ to take action, locate, diagnose, and correct equipment maintenance or operation.

The alternative consulting analyst social organization used by the field study sponsor (Section 2.4) had its own tradeoffs. The consultant analysts maintained knowledge of data analytics and stereotypical actions, while observations and specific actions were the responsibility of the customer ‘energy specialists’. This resulted in issues with misunderstandings of energy models

(Section 2.5.7). I did not observe the distributed, decentralized energy management described in Japanese manufacturing systems (Fawkes, 1986), but briefly speculate on implications of these three social arrangements for M&T software design.

The intermediate steps of knowledge in the Comparative Analysis strategy can be socially divided according to the background shading in Figure 15. If the modeling approach is complex, it is possible that just one analyst (or a machine learning programmer) knows the details of the model training assumptions and dataset (Figure 15 inner dark circle). Software tools sometimes do not allow the raw meter and ‘driver’ data (lower half of Figure 15) to be observed except by instrumentation technologists. This is a barrier to consulting and decentralized M&T approaches. Centralized energy ‘gurus’ can overcome poor M&T tool features, but decentralized non-specialist workers cannot be expected to develop perfect knowledge of these details. They may need to ask system engineers and analysts their intention or understanding (at right, Figure 15). In a consulting social arrangement, this will require communication by an off-site (and possibly off-contract) implementation team, which an M&T tool could support.

**Table 11 - A conceptual mapping of how well three social organizations might support the six strategies presented above. Filled circles represent the author’s judgment of most suitable, empty least suitable.**

<i>Least ○●●Most</i>	<b>Comparative Analysis</b>	<b>Equipment Inventory</b>	<b>Qty. &amp; Cost Analysis</b>	<b>Condition survey</b>	<b>Profile Patterns</b>	<b>Event Association</b>
<i>Centralized Guru</i>	●	○	●	○	○	○
<i>Outside Consultant</i>	●	○	●	○	○	○
<i>Decentralized</i>	○	○	○	●	○	●

Table 11 visualizes how the three social arrangements might interact with strategies. Centralized on-site expert ‘gurus’ can develop analyses, but still depend on colleagues for information about business operation and for help in changing work practices. Outside consultants can specialize in model-building and apply cross-domain experience, but have even less physical presence and associated social influence. Decentralized arrangements with non-specialist workers will be less

able to cultivate data or models, but are present to perform ‘good housekeeping’ surveys, recognize signs of work activity in profile patterns, and recall energy-related events. We conclude by discussing this analysis and its implications for the design and evaluation phases of this dissertation.

### 3.7 Discussion:

The CWA presented above describes prerequisites for control of business energy consumption: a) purposeful opportunities to affect system state (Section 3.3), b) activities to understand and affect system state (Section 3.4), and c) mental and external models used in cognition (Section 3.5). It is intended to qualitatively summarize and structure literature review and M&T observations from interviews, participation, and field study. While not validated, and based on analyst judgment, it can be contrasted with other analyses or extended to support more rigorous application. Next I discuss factors from the analysis relevant to design of a cost-effective diagnosis support tool for M&T. What complexity is necessary? What features of a work domain, what knowledge states, and which strategies might be most productive to support?

#### 3.7.1. Work Domain change and representation maintenance

Half of the M&T strategies I determined explicitly developed representations of work domain state or structure (Table 7). However, M&T work domains are not static, but are instead business systems with an associated pace of change. Within a business, structures (e.g. equipment, processes, other cost tradeoffs) will evolve over time. Sensors and measurement practices may degrade, changing data records’ meaning. Since representations can fall out of date, workers will have to understand and reason about the age and representativeness of data records or models to appropriately rely on strategy output (e.g. whether Comparative Analysis performance indicators are credible). Some strategies are less vulnerable as they use simple enough representations that can be invariant (e.g. Consumption & Cost Analysis normalizing by building size or manufacturing production) or are only expected to represent a snapshot in time (e.g. an energy audit).

An irony is that the more successful energy management tasks, the more they change the work domain. The goals of M&T are to inform capital investment, refine energy consuming

operations, and improve equipment condition (Section 1.1.1). Therefore strategies that depend on representation are penalized by achieving task goals.

### 3.7.2. Including representation in Control Tasks

It could be argued that the work function of cultivating data & models (Figure 14) is a “workaround” (Vicente, 1999, p. 112). In CWA theory a workaround task is a transient artifact of a particular information system and way of work. Just as sensors are not usually included in WDA (as they can be redesigned), workaround tasks are not usually analyzed in ConTA. However in the case of M&T, explicit representations inform other tasks, prove financial business cases, allow historic comparisons, and support statistical analyses. If M&T is fundamentally about diagnosing unobservable structural changes in a work domain, then can an analysis ignore the process of determining a comparator for change? Representation-maintaining work is also useful to distinguish because it is not directly productive and cost-effectiveness is the primary criteria of M&T activities (discussed below in 3.7.3).

Representation-maintaining is essential to many strategies (Table 7). Strategies depend on representation in varying degrees. Some, particularly condition surveys, are more ecological (Gibson, 1979) and can leverage human abilities to directly perceive energy waste without needing to plan or mentally simulate. But for social tasks, representations are how information is communicated between colleagues and over time. Developing and reconciling representations to describe reality are fundamental to Equipment Inventory, Consumption & Cost, and Comparative Analysis strategies. The process of either updating one’s beliefs (e.g. finding and making a note of forgotten-about equipment) or correcting the business system (e.g. unplugging an un-used fridge) may best be characterized as an abductive reasoning process (Bennett & Flach, 2011). Explicitly analyzing belief maintenance as part of a decision ladder notation may be a useful extension to ConTA methods.

### 3.7.3. Representation vs. productive work

Interview and field studies confirmed the prevalence of representation-maintaining work, but also emphasized the pitfall of M&T practitioners over-spending on cultivating data and models (Hilliard et al., 2009). Cultivating data and models is an investment, anticipating that the knowledge base will be useful in problem-solving (or justifying capital investment) later. If no-

one does energy cost-controlling work, or it is ineffective, the time (and information technology) spent cultivating data & models is wasted.

An irony of both ‘Cultivate data and models’ and ‘Control energy costs’ work functions is that time spent performing them harms task performance. This labor tradeoff is not unique (it is particularly of concern in military domains), but it is less important in safety-critical or salaried tasks considered by Human Factors practitioners. Time-efficiency is a key constraint on successful M&T, and M&T tools should discourage workers from prioritizing data or model maintenance over energy control actions.

### 3.7.4. Ambiguity requires knowledge-based behavior

Developing understanding is a difficult-to quantify benefit of performing M&T tasks.

Organizationally, much energy management work focuses on education programs (BRESCU, 2001). Education is less important if monitoring energy performance of a stable, well-behaved group of equipment (e.g. a manufacturing machine), where workers over time will develop habitual rule-based behavioral responses (RBB). But controlling energy is non-routine when influenced by interactions of work domain changes, situation-specific task goals (e.g. Table 6), or particular statistical models. Unfamiliar work is more likely to require deliberate knowledge-based behavior (KBB), particularly for social situations where workers need to discuss and defend their conclusions.

For example, to build up certainty in a belief, workers must consider inference chains from observations to justified actions (Appendix A.5). Depending on organizational structure, workers may need to be very comfortable in the correctness of their beliefs before they take social risks interpreting statistical conclusions. And if they need support from colleagues, they may need to educate them. In any case, KBB is effortful, time-consuming, and error-prone. M&T tools should support opportunities to enable recognition and RBB, but also be inspectable to support knowledge-based reasoning. The less complex the tool, the easier it can be inspected.

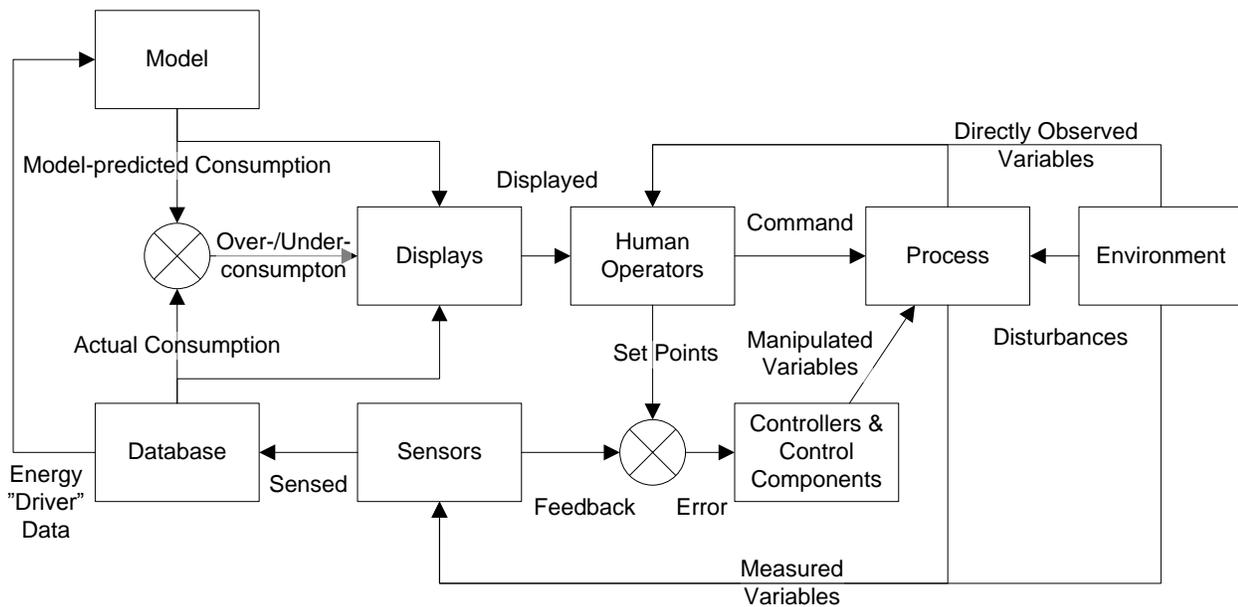
### 3.7.5. Control-Theoretic perspective: sophistication versus ambiguity

Because work domain change means models will fall out of date, and because of the cost penalty and social risk of model maintenance, a crucial choice in M&T tool design is how complex analytic models should be (Section 2.5.5). For the comparative analysis strategy, this can be

considered by exploring the relation between complexity of statistical models and corresponding mental models needed to interpret the result. As a thought experiment, two extremes are:

- No statistical model, where a person's mental model needs to interpret whether energy consumption is good or bad, correcting for every system change (i.e. the profile patterns strategy)
- A Platonic ideal statistical model, which corrects for every productive variation, and whose residual represents pure 'over/underconsumption'

On face value, the second case would seem obviously better. Such a model would allow Comparative Analysis performance indicators to be perfectly reliable and presumably require no understanding from workers. But in M&T, trying to achieve this perfect model may not be effective. I will outline an argument with reference to a control-theoretic illustration shown in Figure 16. Elements shown in Figure 16 are capitalized in the text (as in Jamieson & Vicente, 2005).



**Figure 16 - Comparative Analysis strategy as a feedback control loop for a System subject to Disturbances. (Adapted from Jamieson & Vicente, 2005)**

Drawbacks of more complex models for Comparative Analysis in M&T include:

- Change in the work domain (Process, Sensors, Controllers, Control Components and Environment in Figure 16) may require change in the Model. Which Human Operator maintains

the Model? How does this labor cost compare with the energy costs avoided? What are the implications if the Model is out-of-date (Lee & See, 2004)?

- What counts as Over-/Under-consumption? This depends entirely on the modelers' intent (Section 2.5.5 and Figure 7). An accountant's wasted energy is a superintendent's saved labor costs.
- More complex models rely on more Sensors or Databases. If one fails, how will the analytic model degrade?
- Will Human Operators be able to develop understanding to apply Directly Observed Variables to alternate task strategies?

But the main argument against Model complexity is that it propagates ambiguity in Process and Environment state into the diagnosis problem. Complex sociotechnical systems subjected to environmental Disturbances are always difficult to control. The more complex a Model and the more Measured Variables it (potentially) corrects for, the more complex a mental model is required to evaluate not just the Model output, but all its inputs in Figure 16. Which Disturbances or Measured Variables does the Model-predicted Consumption correct for? Which Disturbances are Measured by Sensors, and is their data still representative? This may require increasing Directly Observed Variables, reducing labor savings. More complex models increase the information required to be evaluated, which risks making the diagnosis problem formally underspecified (Jamieson & Vicente, 2005, p. 15).

### 3.7.6. Limits of analysis

Before concluding the discussion of the work analysis, some caveats should be reiterated. This analysis:

- Included only a generic work domain analysis, not an analysis of a particular system.
- Did not consider the causal structure of work domains in detail
- Mentioned but did not study intentional constraints of financial contracts and energy pricing. For utilities like electricity, their time-value may vary by orders of magnitude.
- Only analyzed strategies for the Control Energy Costs work function (not for developing particular representations)
- Did not include a fine-grained descriptive strategies analysis.

- Was based on literature, interview, and field studies of a limited Canadian context (Section 2.4.4).
- Did not validate the analysis.

These are opportunities for future human factors analyses of M&T work.

## 3.8 Conclusions for design

This chapter concludes with the implications from the Work Analysis that guided design of a work support tool for M&T (Chapter 4), and a subsequent experimental evaluation (Chapter 5). I concluded that M&T tools should support diagnostic reasoning about uncertainty, support diagnostic search in additional work domain structures, and require no more representation-maintaining work than necessary.

### 3.8.1. Support for diagnosis, beyond ‘alerting’

First, the work domain specifies economic pressures on M&T. Work is profitable if marginal energy savings outweigh labor and equipment costs. Therefore M&T tools can succeed either by supporting more effective behavior (including *action*, not just insight), by saving time, or by making work easier to delegate to lower-paid workers.

The Comparative Analysis M&T strategy produces an ‘over/underconsumption’ performance metric that can alert, but contributes to diagnosis only in terms of isolating within time and sub-metering structure. This seems to suffice for management applications (Carbon Trust, 2008; Fawkes, 1988; ISO Technical Committee 242, 2011). Existing Comparative Analysis supports management-by-exception strategies like comparing and ranking energy improvement in a franchised business, or schools within a school board. However, no matter how well management is alerted, orders are not delegation and someone must still search, find the energy-wasting problem, change work practices, and/or perform a repair. Investigation, diagnosis, and problem-solving require skilled workers, are risky, and can consume a lot of time.

### 3.8.2. Making structures searchable

Investigative work usually requires switching strategies. For example, a worker might detect a 240kWh / day overconsumption with Comparative Analysis, cross-check with equipment

inventory for suspect 10kW loads, and perform a Condition Survey of the submetered area looking to see if any of the suspect equipment could mistakenly be operating 24h/day. In coordination, these strategies search different work domain structures (e.g. Table 5). Supporting easier or faster diagnostic searches of the business system is a time-saving opportunity. Considering the more concrete structures described in the WDA (Section 3.3):

- Physical Object structures are searchable through physically walking through the space, as in the Condition Survey strategy. They could be made more searchable with better cameras / sensors to make energy waste more visible (or exposed utility architecture styles)
- Physical Function structures, such as piping, ventilation ducts, wiring networks and so on, can be searchable, such as through the Equipment Inventory strategy. They could be made more searchable with finer-grained sub-meters or ‘smart’ devices. However, those can add model maintenance issues, (discussed above), and are less useful for physical functions that can be flexibly applied to many ends (Table 5).
- Purpose-related functions and processes, such as heating, ventilation, or production processes are harder to search. Two analyzed strategies may offer some support. The Event-action association strategy searches by association with noticeable changes in processes (e.g. more energy consumption because “we filled a lot of orders that day”). The Consumption & Cost Analysis strategy develops performance metrics framed in terms of Purpose-related functions and processes (e.g. kWh/ton production), but only at low-frequency timescales.

The more structures in which a change can be located, the fewer possibilities remain to search. For example, knowing which part of the natural gas piping system is connected to an over-consumption is a start. But knowing which buildings, floors, and rooms are potentially affected, and that the over-consumption is related to the heating process narrows the possibilities greatly.

I conclude that M&T tools should support complimentary strategies that enable different work domain structures to be searched. This will allow workers to apply several strategies in coordination to isolate and diagnose problems. Which work domain structures and what level of detail might be cost-effective?

### 3.8.3. Maximizing benefit of imperfect data and simple models

Model and data structures are part of M&T energy control systems (Figure 16). Since cultivating data and models involves effortful reasoning about ambiguity (Figure 14), benefits *and* costs of models must be considered. Tradeoffs between empirical and cognitive models are an opportunity to optimize design of a “joint cognitive system” (Dalal & Kasper, 1994). M&T tools should not depend on models that are unjustifiably time-intensive to maintain, expertise-demanding to decipher, or harder to assess in diagnosis. Simpler M&T information tools have advantages of fewer sensors, data storage, and processing rules to maintain. They may leave more ambiguity ‘in the world’ to resolve, but at savings of less ambiguity (and sunk cost) in the information support system. I conclude that information system design for M&T should support efficient ‘adaptive laziness’ and wherever possible re-use existing models.

### 3.8.4. What features are needed in M&T software tools?

This discussion has outlined three opportunities for M&T software tools that I pursue through design in Chapter 4. However, this analysis can suggest other opportunities to improve M&T (or energy management) support tools. Designers could consider how their software supports people in:

- Relating existing databases about the business at multiple levels of abstraction (e.g. Table 4)
- Relating sub-meters to decomposed energy-relevant phenomena at multiple levels of abstraction (e.g. Table 5)
- Communicating information between energy management activities (e.g. Table 6)
- Retrieving and processing information within M&T activities (e.g. Figure 14, Appendix A.1)
- Structuring data to enable strategies (Table 7), for example multiple temporal reference frames relevant to Profile Patterns strategy (by-shift, or by-production-run, or by-special-event)

Recently developed (though not widely adopted) enterprise energy management information systems are starting to address some of these opportunities, particularly more abstract work-domain structures (Tanaka, Watanabe, & Endou, 2010). However, there is much room for improvement. The remainder of this dissertation describes a contribution to improving the state of the art in M&T information support.

## Chapter 4

# Applying Work Analysis to develop two novel M&T diagnosis aids

### 4.1 Motivation and Opportunity

After field study and work analysis, I developed several concepts for M&T information system features, prioritized two in coordination with a commercial software developer, and with a research assistant implemented both as functional off-line prototypes. While the design process is not a core contribution of this dissertation, this chapter will briefly introduce the motivation and objectives, describe one of the prototypes in detail, and introduce an experimental evaluation of the prototype in Chapter 5.

#### 4.1.1. Collaboration with software developer

The prototypes were developed in collaboration with Energent Inc., the M&T software company that participated in the field study of Section 2.4. The design intent was to propose extensions to the existing M&T tool, which could be collaboratively implemented by Energent's development team. The vendor was engaged throughout the design process, and helped prioritize two of eight product feature concepts for implementation, as described in this chapter.

#### 4.1.2. Design Objectives

The three design objectives set in Section 3.8 guided the design process. I sought to

- 1) Support diagnosis, in this case by extending the (popular) Comparative Analysis / CUSUM strategy for M&T
- 2) Enable diagnostic search through the Generalized / Purpose-Related Function structure of the work domain, as reflected in energy models, and
- 3) Require no additional model-development or maintenance work compared to existing CUSUM-based methods.

Two additional practical objectives included:

- 4) Develop concepts that could augment the existing M&T software in use by Energent's in-house analysts and clients (Section 2)
- 5) Based on feedback from the field study (Hilliard & Jamieson, 2014b), ensure that concepts could fall back to being useful and usable in the form of paper print-outs. Interactive Web applications like the existing M&T software can navigate large energy data sets, encourage decentralized use, and distribute up-to-date information. However, a design that also functions on paper becomes archival, reliable, easily shared, flexibly annotated, and more approachable for older or non-technical users.

Achieving the diagnosis and search-related objectives while keeping distinct, simple product concepts challenged design methods and suggested two distinct sets of information required.

### 4.1.3. Design Methods

The Cognitive Work Analysis of M&T developed in Chapter 3 was conducted in part to inform the design efforts described here. I expected to apply the Ecological Interface Design (EID) framework whose theoretic principles are (Vicente & Rasmussen, 1992):

- Seek psychologically relevant regularities in the work environment (e.g. functional structure)
- Design an interface that unambiguously represents these useful regularities (e.g. with analogies or metaphors)
- Implement interfaces (not just displays) wherever possible, to support workers in thought experiments and actively interrogating the world to abductively maintain their mental models (Bennett & Flach, 2011).
- Format the interface so that workers can perceive/control it with any mode of cognitive control.

As I investigated M&T it became apparent that standard EID methods (Burns & Hajdukiewicz, 2004) were not well-suited to the M&T domain (Hilliard & Jamieson, 2014a). In standard, well-developed EID methods, Work Domain Analysis (WDA) serves as a psychologically relevant framework to represent system regularities in terms of structural means-ends, part-whole, and causal/topographical relationships (Rasmussen et al., 1994). This analysis then specifies information content requirements and informs the design of analogical graphic forms whose surface features behave consistently with deep regularities of system function (Rasmussen, 1974).

However, as discussed in Section 3.3, it is less useful to analyze a particular work domain, since software tools must be general-purpose to be cost-effective. Fortunately, WDA is not the only system structure model that was proposed for EID. Rasmussen suggested “support should not aim at a particular process, but at an effective *strategy*, i.e. a category of processes.” (Rasmussen & Vicente, 1989, p. 525) and later explicitly recommended that “systems serving an autonomous user ... [where] the intentionality depends entirely on the users’ needs” (Rasmussen et al., 1994, p. 187), should be designed with *cognitive strategies* as the basis for an ecological interface. I did not pursue a formal EID design process, but instead sought to satisfy the principles of EID based on the information required (or developed) by comparative analysis strategies.

#### 4.1.4. Information Requirements

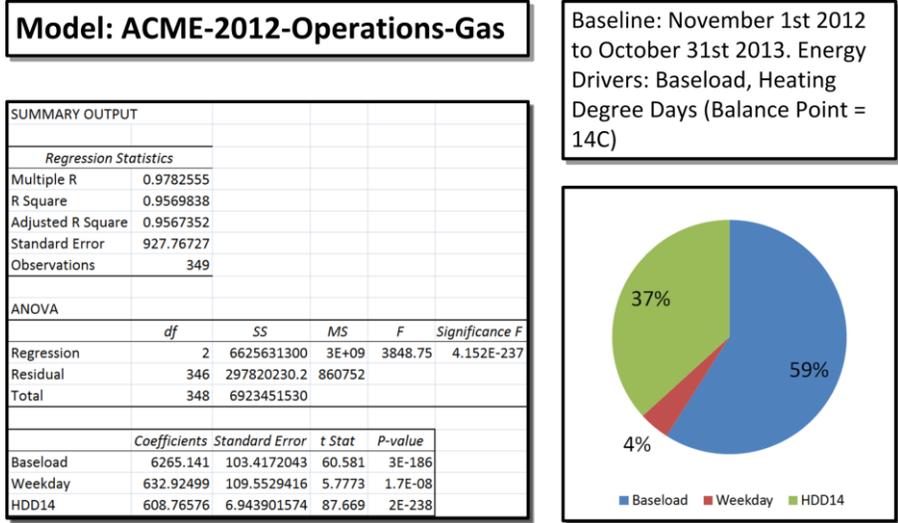
The task analysis of Section 3.4.2 and the Comparative Analysis strategy analysis in Section 3.5.4 served as a coarse set of information requirements for the prototype software design. The analyses describe knowledge that workers will need to develop at some point during work. Of course this does not define which data and information should be prioritized or how it should be presented. Pre-analysis of priorities is complicated by M&T being applied in a variety of business system.

The approach I followed was to adapt a new statistical strategy to support diagnosis within Comparative Analysis (introduced in Section 3.5.4). This new strategy supports searching within work domain structure captured by the parameters of the energy model (Sections 2.3.3 and 3.5.4). This introduces an information requirement of understanding model parameters. For someone to be able to correctly interpret what a change in a model parameter means, they need information about what exactly the model represents. I developed a companion report to support assessing ambiguity in the Comparative Analysis model (Section 4.2), and discuss each in turn.

## 4.2 Prototype: Model Summary Sheet

The Model Summary Sheet was designed to supplement existing information sources about energy model validity (Figure 17) that field study participants found insufficient to assess the meaning of CUSUM charts. Since the novel diagnosis aid described later in Section 4.3 aims to support diagnostic search, clear information mapping model parameters to real work system features is essential. The final Model Summary Sheet prototype is shown in Figure 18 and

printed elsewhere in full-color (Hilliard et al., 2014). While it is intended to be implemented in parallel with the diagnostic aid described next, I did not experimentally evaluate the Model Summary Sheet and do not emphasize it as a contribution in this dissertation.



**Figure 17 - Examples of how linear regression models used for CUSUM comparative analysis were presented to end-users. This is representative of three M&T software systems I observed in the course of participant observation and field study.**

CLIENT LOGO

VENDOR LOGO

## Energy Model Summary Report

**Acme Hospital**  
*Natural Gas*

**Main Building**  
*Baseline Model*

Standards Compliance\*

ASHRAE 14:2002 ✓

IPMVP Option C ✓

**Model Executive Summary:**

*This energy model captures Acme Hospital's natural gas performance at 2009-2010 levels, adjusted for weather and week-ends. It can be used both for day-to-day operations and to track effects of building insulation retrofits in the Baker Wing.*

*Jane Doe*

**Model Update #2 of 3**

Trained from *November 1<sup>st</sup>, 2009* to *Oct. 31, 2010*

In Service *November 1<sup>st</sup>, 2010* to *May 1<sup>st</sup>, 2012*

Consumption over Training Period

**3,705,309 m3 gas**     4% Internal

**\$ 741,061 / year**     37% Weather

Annualized cost estimate, **\$0.20 / m3 gas**     59% Baseload

**Site ESCO**

Model Training Summary

Fit *November 1<sup>st</sup>, 2009* to *Oct. 31, 2010*  
to *Natural Gas* meter *Revenue-1-E615*

349 data points over 365 days

16 missing (4.4%)  
0 excluded (0%)  
16 eliminated (4.4%)

Training Variance → 0%

**Acme Hospital**

Performance Summary

12 Months training data ✓

100% Operation modes covered ★★

100% High / Lo range covered ★★

4.4% Training data eliminated ★★

8.7% CVRMSE

ASHRAE 14:2002 / IPMVP Option C

**Baseline**

Model Interpretation Guide:

*Data excluded because:* Three periods of missing data in December 2009, January, and February 2010. Because of this, the model may slightly under-estimate heating.

8.7% typical day-to-day variance

96% of day-to-day energy variation explained

CVRMSE: Highly Variable

R<sup>2</sup>: Long-term Average

Consistently Accurate

Very Responsive

**Consumption & Model**

**Site ESCO**

Model Driver Breakdown

4.2% Weekday

37% HDD14

59% Baseload

**Baseline**

Driver Parameter

633 m3gas/weekday

608 m3gas/degree C

6265 m3gas/day

**Acme Hospital**

Trained Range\*

1 Week-0 Day

29 degrees 0 C

**Model Driver Descriptions:**

**Weekday**  
The hospital uses more hot water during week-days when patient load and procedures scheduled are greater.

**HDD14**  
Heating Degree Days are a measure of how cold the weather is and how much Gas is needed for space heating. For every degree of daily average temperature under 14 Celsius, the site uses proportionately more gas. Boiler #1...4 operation may complicate matters.

**Baseload**  
Daily gas is used mostly for hot water and sterilization.

**Extended Energy Analyst Note:**

>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent ornare leo sed nunc iaculis vitae tempus dui placerat. Quisque ut lacus justo. Praesent sodales, nisi non dapibus aliquet, risus suscipit nisi, sit amet tempus dolor elit, vel quam. Praesent dignibus veherra purus vitae aliquam. Quisque nulla arcu. Interdum nec curcum nunc, semper eget eros. Duis in magna ante, ac hendrerit est. Nunc in blandit nulla. Nulla sollicitudin dolor vitae libero dignissim aliquam. Proin sem lorem, suscipit vitae semper non, feugiat eget felis.

\* This report indicates model training criteria. However, compliance for a specific purpose also depends on other criteria. Training data with the model is not a guarantee of model performance. Ask your analyst about ASHRAE & IPMVP compliance checks for your specific use. ( TO BE CHECKED WITH LEGAL )

**Figure 18 - Model Summary Sheet Prototype, as delivered to client for implementation. First two pages are fixed-format, third page can expand as necessary to accommodate models with more driver variables and associated parameters.**

## 4.3 Prototype: Recursive Parameter Estimates Charts

The second prototype is a time-series diagnosis aid, which augments CUSUM charts without requiring more model development or maintenance. This prototype was developed based on the Recursive Estimates (RE) algorithm (Ploberger et al., 1989; Zeileis, Leisch, Hornik, & Kleiber, 2002), and extends prior applications of linear regression to diagnosis in M&V (Kissock & Eger, 2008). I modified the RE algorithm to reduce ambiguity and better suit the M&T task environment, implemented a prototype, and experimentally evaluated it (Chapter 5). The development of the prototype is explained briefly below and in detail elsewhere (Hilliard & Jamieson, 2013, 2014a).

### 4.3.1. Statistical time-series change detection

The statistical processes used in Comparative Analysis M&T strategies (Section 2.3 and 3.5.4) can also be characterized as 'sequential change-point detection' (Khodadadi & Asgharian, 2008; Krämer, Ploberger, & Alt, 1988), in econometrics. CUSUM is not the only algorithm used for such problems. CUSUM is simple and robust to measurement errors (Young, 2011, p. 42), but responds inconsistently to changes associated with model parameters (as outlined in Section 2.3.6 and 2.5.6).

Another algorithm for structural change detection is Recursive Estimates (RE) or the “fluctuation test” (Ploberger et al., 1989). RE charts use model error to estimate *how changes are explained by model parameters*. This is obviously important in control engineering (Young, 2011), and similarly in M&T applications could inform energy performance diagnosis in terms of processes represented by driver variables (e.g. heating efficiency). However, RE has not been applied to M&T before. The next sections describe the RE algorithm and its potential to support the principles identified in Section 3.8.4 for improving M&T tools. The discussion is structured similarly to the discussion in Section 2.3 of M&T with Recursive Cumulative Sum of Residuals (CUSUM).

### 4.3.2. Origin of Recursive Parameter Estimates Method

The recursive estimates (RE) class of structural change detection methods were invented after energy M&T was developed (Fawkes, 1988; Ploberger et al., 1989). RE methods have since been refined (Kuan & Chen, 1994), implemented in statistical software (Zeileis, 2003), and become a

common econometric analysis tool. Similar to CUSUM charts, RE charts detect system changes based on model residual error, producing a time series empirical fluctuation process. RE and CUSUM “are thus based on identical ingredients, which are however put to different uses. In fact, the [recursive estimates] Fluctuation test can be viewed as a backward CUSUM type procedure.” (Ploberger et al., 1989, p. 312).

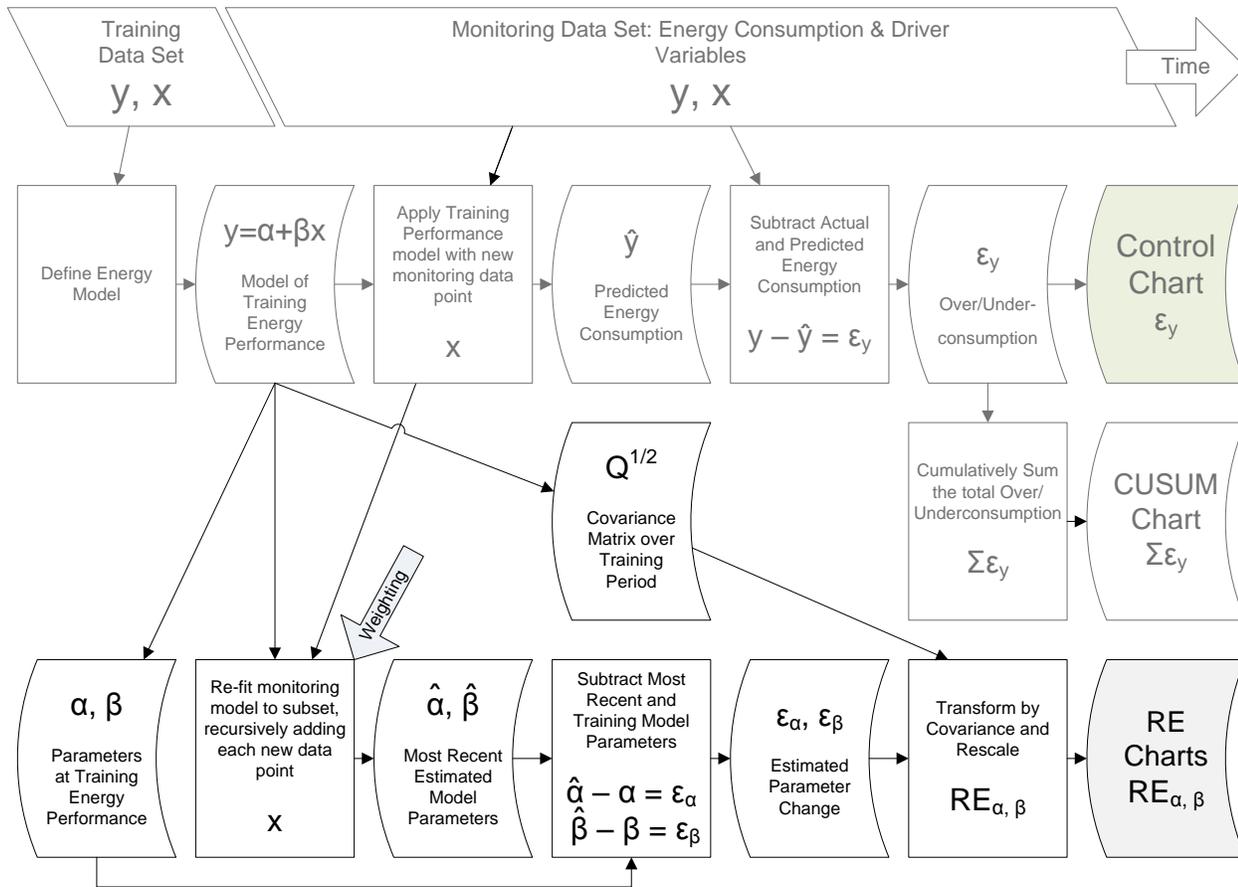
### 4.3.3. Method Outline

RE charts can be applied to M&T through a similar Comparative Analysis strategy as CUSUM charts (Section 2.3.2). Figure 19 outlines the data processing steps of applying RE charts:

- 1) Train a linear regression baseline model, as for a CUSUM approach. Store two model training properties: the ‘historic’ model coefficients  $\beta_n$ , and the normalized covariance matrix  $Q^{1/2}$  of the driver variables over the training period  $X_n$  (Ploberger et al., 1989):

$$- \frac{\text{---}}{\text{---}}$$

- 2) For each time-series sample  $i$  following the training period  $n$ , append the new data to a monitoring set and re-fit the linear model to all data up to time  $i$ . Calculate the difference between recursively fit monitoring model parameters and historic values ( $\beta_{ni} - \beta_n$ ).
- 3) Transform each resulting parameter change at time  $i$  into an RE chart point through vector multiplying by the historic covariance matrix  $Q^{1/2}$ . This corrects for normal inter-parameter coupling present in the training data. Multiplication by  $Q^{1/2}$  resembles Principal Components Analysis, a coordinate transform from the ‘state space’ defined by the driver variables into a new ‘change detection space’ that is more orthogonal to (and thus less affected by) usual covariance between driver variables.
- 4) Normalize the scale of the RE charts by dividing by each parameter’s training variance and a ratio of the monitoring time elapsed to the length of the training period. This produces unit-less values compatible with a hypothesis-testing statistic (Kuan & Chen, 1994).
- 5) Interpret the resulting RE charts. Inflection points mark times where a performance change may have occurred. If using test statistics, note if/when the RE chart exceeds the test statistic limits, and conclude a change has happened in that model parameter.
- 6) For suspected parameter changes, consider what the model parameter represents in the real energy-consuming system and infer where to search for related causes.



**Figure 19 – Data flow diagram for calculating RE charts. Existing CUSUM chart data flow (Figure 4) shown in light grey.**

Similar to CUSUM charts, most steps are algorithmic and more completely described elsewhere (Khodadadi & Asgharian, 2008; Ploberger et al., 1989; Young, 2011; Zeileis, 2003). Judgment in interpreting RE charts is less well-understood. We will next discuss weaknesses in the RE algorithm that lead to statistical and perceptual ambiguity, and some modifications to the standard algorithm that should support simpler judgment rules for M&T use.

#### 4.3.4. Statistical weaknesses in Recursive Estimates Charts

RE charts, moreso than CUSUM, depend on good models trained with driver variables that are independent, timely, accurate, and complete (as recommended by e.g. Efficiency Valuation Organization, 2012). Of course RE methods cannot outperform the data from which the model is derived. If a system change is ambiguously reflected in driver variables, the parameter estimate fluctuations will be ambiguous. Some weaknesses specific to RE are:

- Autocorrelation in residuals, caused for example by time delay between when energy is consumed and when driver variables are measured, distorts RE chart scale (Kuan & Chen, 1994)
- Un-modeled aspects of the energy-consuming system, such as measurement error or missing driver variables, affect RE charts more than CUSUMs (Young, 2011, p. 42). This endogeneity has three main effects:
  1. Regression Dilution: Measurement noise in driver variables will distort parameter estimates, usually underestimating parameters and overestimating baseload.
  2. Regression Contamination: RE charts are not independent, even when transformed by  $Q^{1/2}$  into more orthogonal components. Measurement error, effects of un-modeled drivers, and even single parameter changes will contaminate parameter estimates for other variables according to their covariance.
  3. Test statistic inflation: Changes during training obscure later detection (Perron, 2006).

These flaws in RE are either inherited from linear regression or encountered in applying statistical tests. CUSUM charts are more robust against these flaws and therefore more reliable for detecting changes because they are based on model prediction  $\hat{y}$  (not parameter estimates). However, even if statistically imperfect, RE charts can still be useful to inform diagnosis if their time-series shapes are informative cues to energy-relevant phenomena.

#### 4.3.5. Perceptual ambiguities in Recursive Estimates Charts

For some situations, RE charts in their standard implementation (Zeileis et al., 2002) create time-series shapes with ambiguous perceptual properties that can limit or complicate interpretation.

- Standard RE algorithms recursively grow monitoring sets with each new data point, so the longer they are used, the more sluggish charts behave and the less they distinguish intermittent and persistent changes.
- Constant energy performance does not necessarily produce straight-line RE charts, due to chart scaling for test statistics. Transferring CUSUM interpretation rules (Section 2.3.5) could lead to false alarms.
- Straight-line RE charts can be produced by a driver variable not varying for long periods (such as a winter heating driver that is zero all spring and summer). This means that if a physical system change occurs while a driver is ‘stagnant’ (e.g. someone damages the heater in summertime), the

effect will not be charted until the associated driver varies again. Charts that conflate ‘no change’ and ‘no information’ conceal ambiguity about the date of a change.

- Normalization and coordinate transforms in RE abandon engineering units of regression models, so unlike CUSUM charts, RE axis values do not have direct physical significance.

These challenges to interpreting RE charts exist in part because the algorithm has been developed to match the test statistic, and the test statistic is defined in terms of statistical certainty over accumulated evidence. But, as discussed next, test statistics may not be especially useful in an M&T task environment. Modifying RE to remove the test statistic may allow the charts to be more easily interpreted in an M&T context and better satisfy the principles of Ecological Interface Design (Section 4.1.3).

#### 4.3.6. Utility of test statistics in Recursive Estimates Charts

In M&T, ease of interpretation by informed colleagues is more important than statistical certainty. Omitting test statistics from RE charts allows design freedom to reduce some perceptual ambiguities. To a statistician, discarding formal tests may seem foolish. Test statistics are necessary when risky decisions must be made solely in terms of quantitative data. However, as discussed in Appendix B.1, there are many reasons tests are of limited use in an M&T work environment:

- The pace of change in most businesses will tend to produce multiple overlapping changes, some of which are non-actionable (e.g. a new building renovation, or change in product formulation). Test statistics triggered by known changes are obvious and unhelpful.
- Test statistics must be matched to a hypothesis, which requires not only anticipating a question to be answered, but also communicating this question to workers interpreting charts.
- Machine-readable driver data is limited, so incorporating or correcting for known changes takes time and adds cost. Test statistics cannot be tailored to every hypothesis.
- Statistical assumptions may not hold, making test statistics misleading (because of endogeneity such as measurement error, missing variables, etc.).
- Human-readable or non-modelable data is often plentiful. ‘Hunches’, if informative, are sufficient for colleagues to investigate and gain knowledge on which to base a decision (Klein, 1993).

- Test statistic calculations and interpretations must be explained to colleagues, and may not help to inform or convince them.
- Calculations which are difficult to perform with commodity office tools (e.g. spreadsheets) may not be widely adopted or included in training.

Such objections may explain why existing statistical tests for CUSUM methods (Zeileis, 2003) are not used in M&T practice (Appendix B.1).

### 4.3.7. Disambiguating Recursive Estimates Charts

Removing test statistics will not fix statistical flaws of the RE method, but it allowed three key changes to task-relevant perceptual features of RE charts. The three changes described below modified the perceptual features of RE charts to make them behave more consistently, reduce ambiguity, and give them some (limited) context (according to the principles outlined in Section 4.1.3). The intent is to support more consistent interpretation using rules, without precluding knowledge-based reasoning. Before-and-after chart appearance is illustrated in Figure 20 and Figure 21. Design justifications are summarized below and elsewhere (Hilliard & Jamieson, 2014a).

#### 4.3.7.1. Exponentially Weighted Memory

Conventional RE charts (Figure 20) recursively add new data to an ever-growing monitoring set. This causes the RE chart to behave differently over time (as more data is recursively added), and filters over characteristic time-series shapes of intermittent changes. To make RE charts consistent and time-invariant the monitoring dataset can be time-weighted by an exponential decay. This is normal practice for “badly defined’ systems such as those encountered in ... socio-economic systems analysis” (Young, 2011, p. 49). Exponentially weighting data in the recursive monitoring model (Figure 19) makes parameter estimates respond more to fresh samples than old data. The decay rate can be tuned to represent phenomena of interest, analogous to adjusting a low-pass filter. An optimal decay rate is difficult to pre-specify (Young, 2011), but can instead be interactive, or plot short and long “memory” RE charts simultaneously (Figure 21).

### Monitoring with RE test (recursive estimates test)

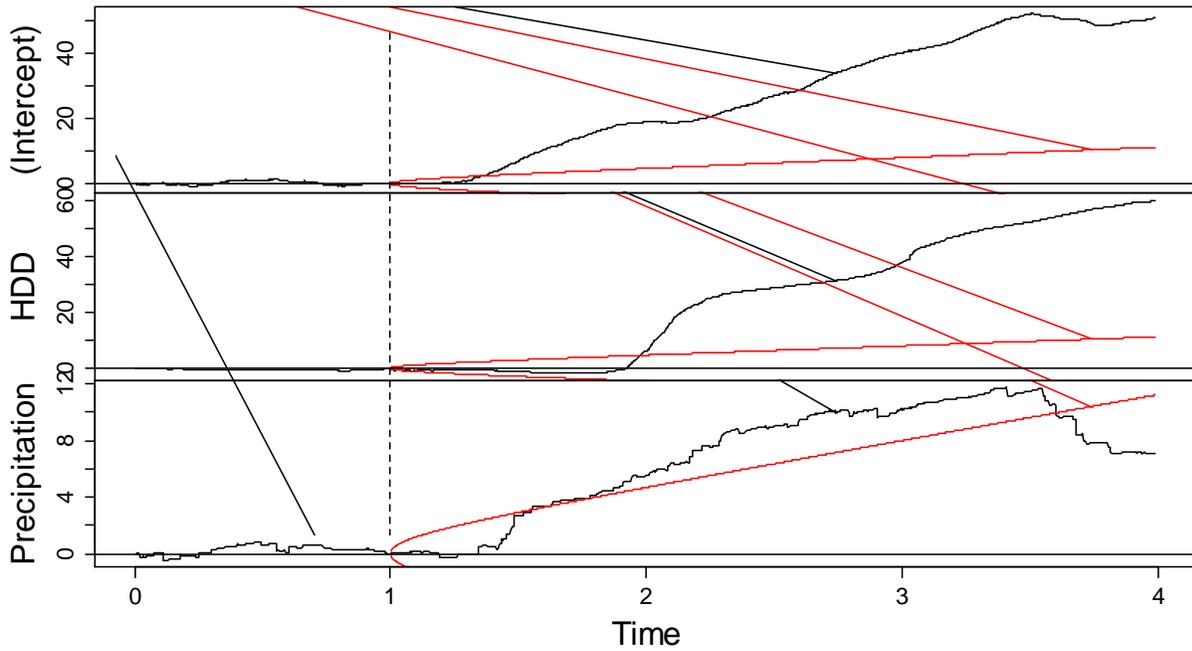


Figure 20 - Synthetic data for a three-term model: Intercept, Heating Degree Days, and Precipitation. Three changes introduced, first to Intercept, second to HDD, and last to Precipitation. RE plots scaled to indicate statistical significance vs. test statistic (red). Contrast with modified RE chart below in Figure 21.

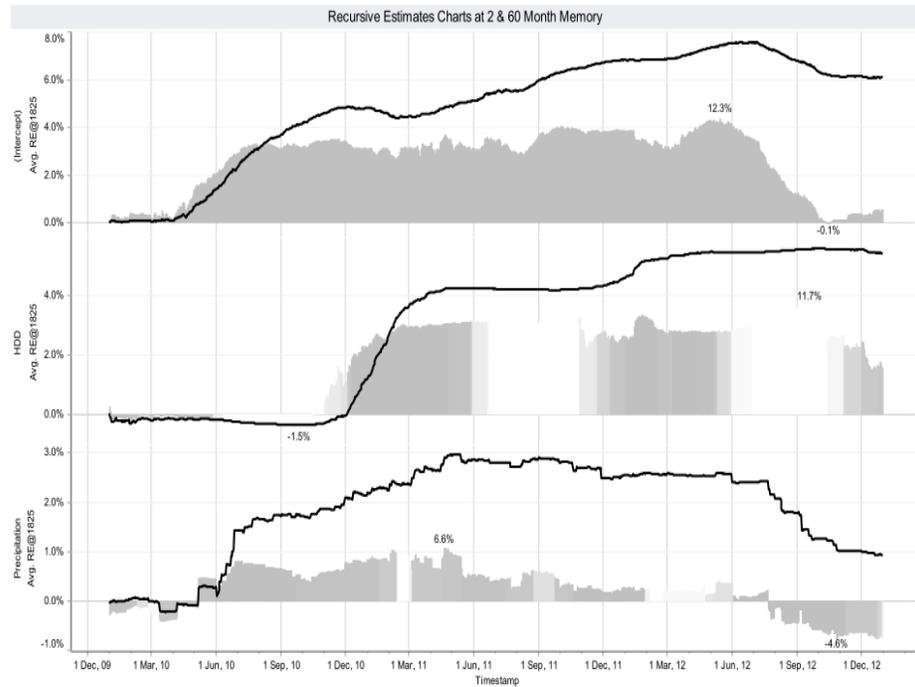


Figure 21 – Same data as Figure 20 in modified RE plots with exponential decay at 60-month and 2-month time constants. Plots scaled analogously to change time and size. The steady height of the middle 2-month (grey) HDD chart shows that the change in heating process efficiency that occurred December 10<sup>th</sup> persisted over two summers with a steady, comparable effect.

Figure 21 shows both 2-month and 5-year time constant RE charts. They are formatted differently to distinguish them and indicate their behavior. The 5-year RE chart resembles and behaves similarly to the CUSUM chart, tracking persistent shifts in performance. The 2-month decay chart resembles and behaves similarly to the Control Chart, depicting abrupt changes as well as uncontrollable variation. Together, these charts disambiguate intermittent and persistent changes, and behave more analogously to underlying statistical shifts. Both these properties should support development of rule-based behavior.

#### 4.3.7.2. Meta-Information

The second change addresses intermittent drivers common in M&T models. For example, seasonal heating/cooling factors are inactive half the year and individual products may be manufactured only from time-to-time. While a driver is constant (zero, for example), it conveys no information about its influence on energy performance (e.g. you cannot tell the furnace is broken until you turn it on). Therefore during such periods, an RE chart should be de-emphasized since it is not informative.

A simple proxy metric for the informativeness of an RE chart point  $\mathbf{I}_{x,n}$  is relative driver variation, which can be calculated as the ratio of variation in ‘recent’ monitoring data to that during model training. I defined ‘recent’ as within one exponential decay time constant (as used for weighting in Section 4.3.7.1).

---

Where  $x_i$  is the value of driver variable  $n$  at monitoring time  $i$ ,  $T_e$  is the exponential decay time constant, and the model training period is from time  $i = 0$  to  $1$ . The relative driver variation metric is mapped in Figure 21 to the RE chart’s transparency. This disambiguates ‘no information’ and ‘no change’, by fading an RE chart into the background as the parameter estimate ages. Mapping transparency to ‘informativeness’ can also show gaps caused by missing data.

### 4.3.7.3. Relative Scaling

A final RE chart design choice is axis units. Conventional RE charts (Figure 20) normalize each parameter according to historic variability and present each on dimensionless test statistic axes. By contrast, a practical benefit of CUSUM charts is that their Y-axis has clear practical significance as “accumulated over/underconsumption” that can be used as a quantitative performance metric.

Only *relative* change in estimated parameters is meaningful, as statistical properties of the RE algorithm prevent determining precise parameter estimates. Therefore axis units should be chosen to encourage qualitative exploratory diagnosis rather than quantitative M&V use. I chose to normalize each parameter deviation  $\hat{\beta}_n - \beta_n$  by its historic trained model parameter  $\beta_n$  in  $Q^{1/2}$  coordinates.

$$\frac{\hat{\beta}_n - \beta_n}{\beta_n Q^{1/2}}$$

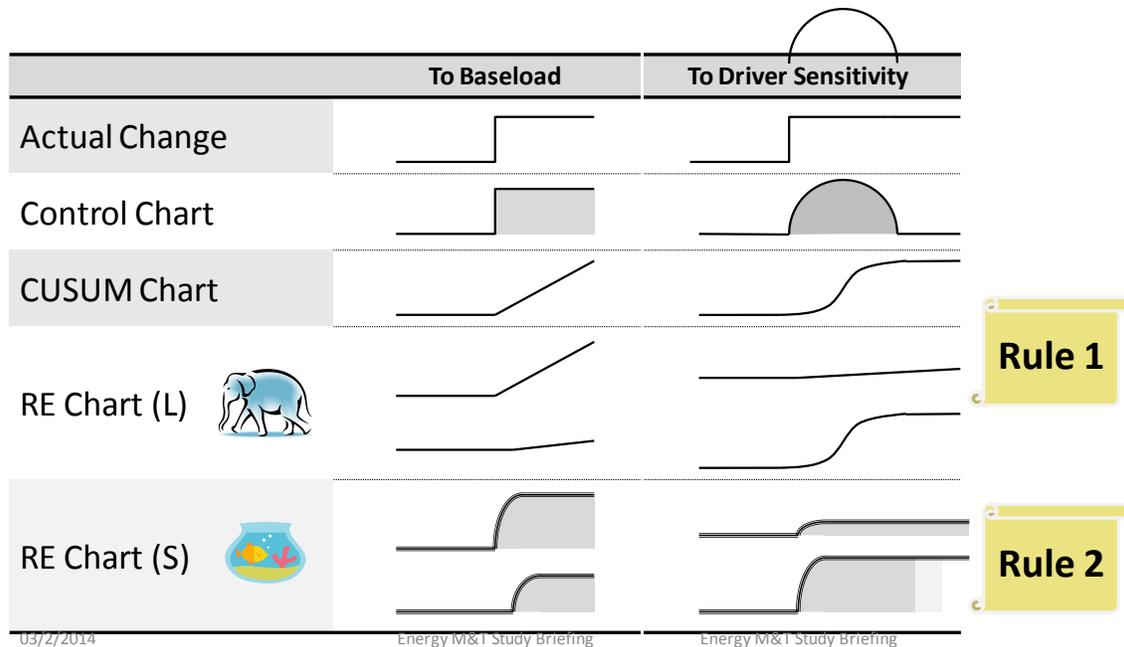
This creates an RE chart with “percentage change” units, which while still without direct physical significance at least indicates the order of magnitude of a possible change<sup>7</sup>, possibly helping to narrow the search scope.

### 4.3.8. Interpreting Modified RE Charts

Specifying interpretation rules for a new statistical application is a bit imprudent, as the interactions of different business work domains, data collection, modeling, and disturbances will produce hard-to anticipate patterns. Consistent with EID principles, the design intent is to produce a representation with enough structure to allow local experts to develop effective rule-based behaviors. However for instructing novices two rules seem a good starting point and are illustrated below in Figure 22.

---

<sup>7</sup> The more orthogonal the independent driver variables used in the energy performance model, the closer the modified RE chart axes will mirror the regression estimated parameter changes



**Figure 22 - Stereotypical behavior of RE charts at long (L) and short (S) exponentially decayed memory, compared to CUSUM charts and the actual underlying change reflected in model intercept (baseload), or a parameter (driver sensitivity). Interpretation rules 1 and 2 apply consistently to baseload and driver changes.**

The first step is to use a CUSUM chart to identify suspected change times, as described in Section 2.3.5. CUSUM charts are the most reliable indicator of practically significant changes. Ignoring known or intentional changes, two tactical rules for each unexplained change are:

- 1) Check for RE chart changes that are *simultaneous and same-direction* as the CUSUM chart.
  - a Compare long memory RE charts with the CUSUM chart to see which parameter's chart "most resembles" the CUSUM at the time of the suspected change.
  - b Use short memory RE charts to confirm the same parameters' chart shifted at the same time.
- 2) If one parameter's RE chart seems consistent with the CUSUM change, use short memory RE charts to determine whether the change persisted or reverted.
  - a If the change persists before or after, note other CUSUM changes that occur without shifts in the short-memory RE charts. These may be due to the same root cause (such as poor performance in rarely-used equipment). Include extra dates in subsequent search.
  - b If the change reverted, note the date it reverted, ensure it is consistent with the CUSUM, and investigate historic events consistent with the start/stop dates.

These interpretation rules can be supported by perceptual aids.

### 4.3.9. Annotating Modified RE Charts

The two interpretation rules above are only valid when comparing the CUSUM chart with RE charts at the same point in time. The reason for this is that RE charts are less statistically robust than CUSUM charts (Section 4.3.4) to measurement error, autocorrelation, or unmodeled driver variables. Even with high quality models, changes will cause movement in all RE charts since RE charts are scaled by the covariance of the model data over the training period (Section 4.3.3). Finally, real-world changes can be reflected in more than one model parameter. For these reasons, RE charts *at the time of a change in the CUSUM charts* are the most diagnostic.

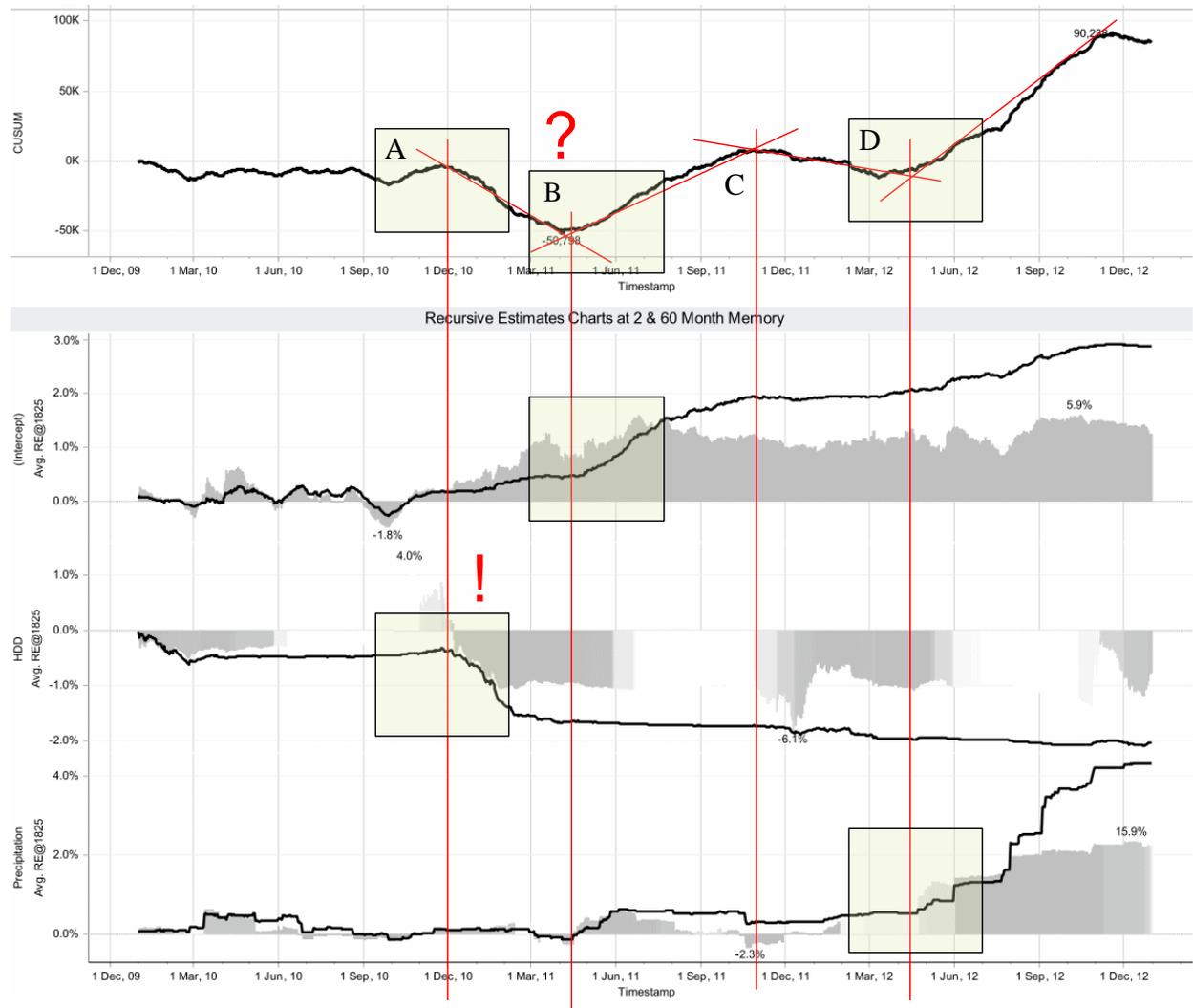
CUSUM and RE charts can be perceptually linked by a vertical line at a time of suspected system change. If the charts are used in paper format, lines can be drawn with pencil and ruler, or by applying a paper mask. This aids applying Rule 1 and Rule 2 (above) and in distinguishing diagnostic from un-diagnostic chart changes. A case study will demonstrate applying these rules to simulated and real data.

## 4.4 Example Recursive Estimates Charts

Two examples are presented below to illustrate how the modified RE charts appear compared to existing time-series charts for M&T.

### 4.4.1. Synthetic data

A common CUSUM chart ambiguity observed during the field study occurs when effects of multiple superimposed changes interacting create a difficult-to-interpret CUSUM chart (illustrated with synthetic data in Figure 9 of Section 2.5). This ambiguity occurs more often in business environments with a fast pace of change, where non-actionable changes can persist and accumulate. For example, if two counteracting changes overlap, such as a reduction in a driver-specific energy use (e.g. improved insulation) and an increase in an everyday energy consumption (e.g. hot water leak), the CUSUM chart's net energy savings or loss will vary seasonally. From inspecting Figure 9, it is not clear how many changes may have occurred. The same synthetic data is presented as an RE chart below in Figure 23, annotated to match a description of how the chart can be interpreted using the rules described above in Section 4.3.8.



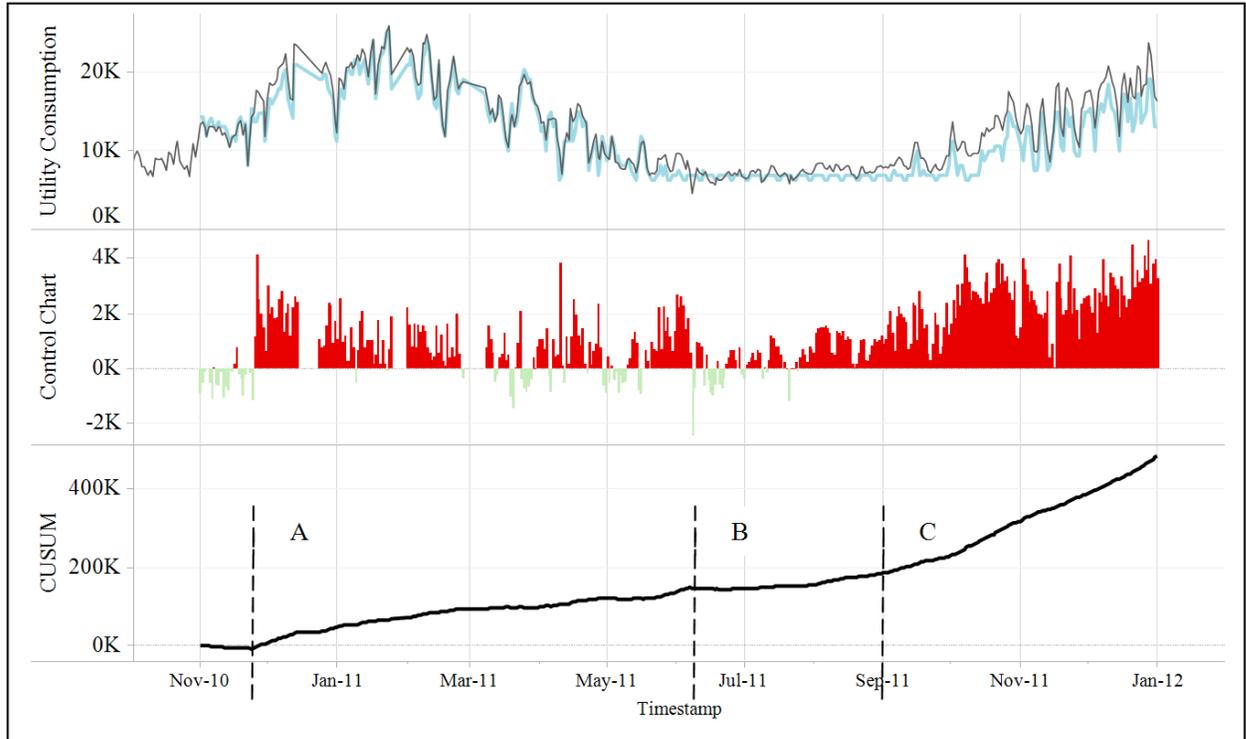
**Figure 23 - Example of CUSUM chart from Figure 9, with RE charts developed from same model, and a superimposed “solution”. The linear model has an intercept (baseload) plus two parameters, seasonal heating plus intermittent precipitation. RE charts are plotted at two exponential decay timescales: 2-year (black line) and 2-month (grey shaded).**

Applying Rule 1 as described in Section 4.3.8, the first change in the CUSUM chart of Figure 23, labeled “A”, occurs simultaneously with a downward shift in the middle RE chart (representing HDD, Heating Degree Days). Similarly, the second CUSUM slope change B coincides with an upwards shift in the top RE chart (representing Intercept, or everyday Base-load). The third apparent shift C will be discussed below. Finally, the fourth shift D coincides with a shift in the bottom RE chart. These are all examples of interpretation that could be trained into rule-based expert behaviors.

Rule 2 suggests that shift C in the CUSUM is a ‘phantom’ change. The CUSUM chart shift occurs without any simultaneous shift in RE charts, whose short-memory (grey) charts hold steady. This CUSUM chart shift indicates that savings to heating efficiency (change A) are affecting system performance again in the subsequent winter, counteracting increases in everyday consumption (change B). This synthetic data was developed to illustrate this example. Next, we show an example with real data.

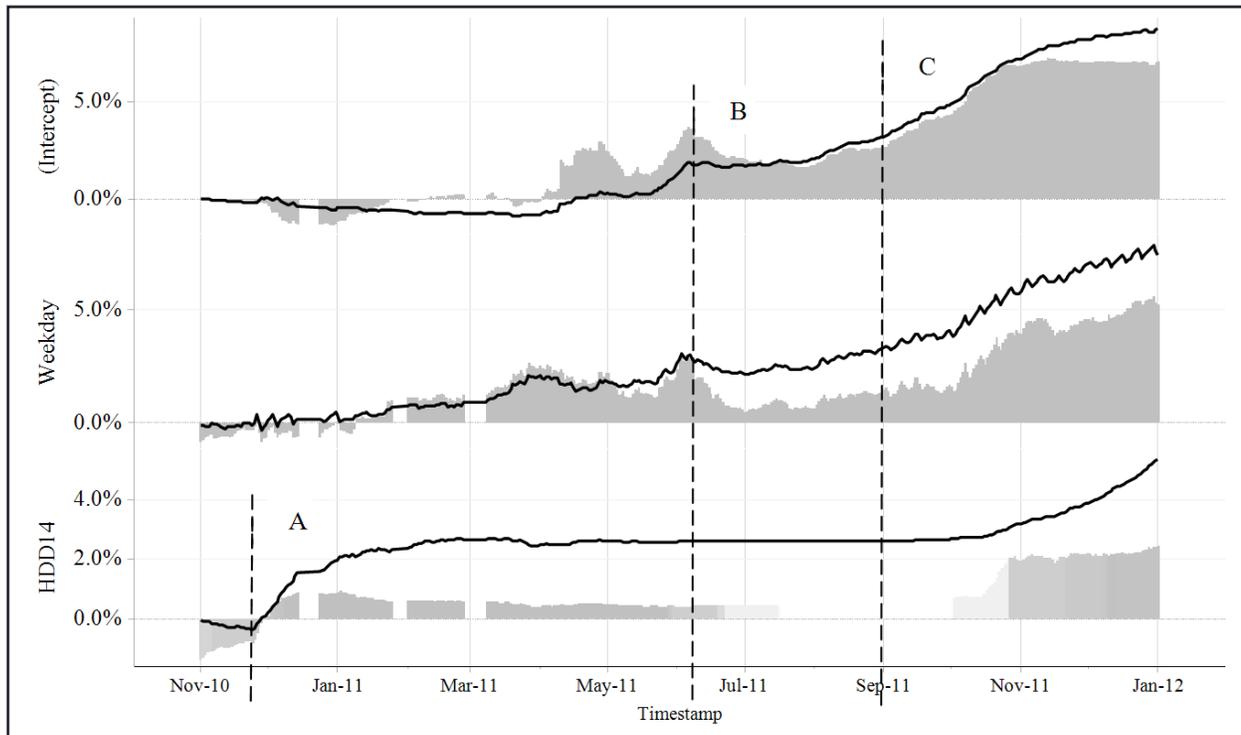
#### 4.4.2. Hospital example

The modified RE method can also be demonstrated on actual data from the hospital which later participated in the field study of Section 2.4. The base truths of the changes discussed in this case study have been confirmed, but as with any real utility data, other un-investigated changes are also present. The Natural Gas consumption model used for this exercise was trained on 350 data points over 1 year, from November 2009-2010, correcting for Weekday/Weekend variation and Heating demand. It is fairly high quality (ASHRAE Guideline Project Committee 14P, 2002), with a  $R^2$  of 0.96, and a CVRMSE of 8.7%.



**Figure 24 - CUSUM Chart, Healthcare Institution Natural Gas Consumption, Fall 2010-2011. From top to bottom: Consumption (thin black) and Modeled consumption (thick blue) charts. Control Chart showing over/underconsumption. Finally, CUSUM chart showing times of three suspected changes (A,B,C)**

Interpreting the site CUSUM chart in Figure 24 using the rules described in Section 2.3.5 might suggest three changes – One in late November 2010 (A), which moderates by June 2011 (B), and a second, larger change around September 2011 (C). The reader can try diagnosis using the CUSUM chart before examining Figure 25.



**Figure 25 - RE Chart, Hospital Natural Gas Consumption, Fall 2010-2011.** RE charts for the same model and period as Figure 1. Each chart indicates change in a model parameter: baseload (Intercept), Weekday, and Heating Degree Day. Each chart shows two exponential memory decay factors: Grey bars indicate a fast-responding, 2-month time constant. Black lines indicate a slow-responding 60-month time constant.

The three changes identified from the CUSUM chart are marked on Figure 25. ‘Change A’ corresponds to an abrupt change in the Heating Degree Day (HDD) parameter on November 24<sup>th</sup>. The short-memory RE chart shows that the heating sensitivity moderated after January 2011, but did not completely revert by summer (coincident with the flattening of the CUSUM slope at time ‘B’). The abrupt HDD increase at ‘A’ is consistent with known events at the hospital: a contagious disease outbreak required the heating system be changed to 100% fresh air, temporarily increasing the building’s sensitivity to cold weather. It is possible that the heating controls were not reverted completely after the outbreak was contained.

At the time of Change ‘B’, Figure 25’s charts of baseload (Intercept) and Weekday parameters had been increasing, indicating gas consumption unrelated to outdoor temperature. Both reduce slightly as the (stable) heating parameter “fades out” during the summer. This change may indicate known inefficient boiler operation at the end of the heating season. Note that the brief “blips” are more noticeable on the RE charts than the CUSUM chart.

Change ‘C’ can be distinguished as three separate events – an increase in the Baseload intercept over September-October, followed by increased Weekday consumption, and finally an increase in the Heating parameter when its driver varies again in late October. This is consistent with the commissioning of a new building wing, whose interior was being finished during the fall. The unoccupied space may have required heating at a higher temperature break-point in October, but over the winter the increased heating load seems quite stable. This example demonstrates how RE charts can help diagnose CUSUM changes. These changes are only those that could be verified after-the-fact, but discovering unplanned changes on-line follows similar rules, with subsequent search in real-time data and by colleagues.

## 4.5 Conclusion: Implementation and Evaluation

Both the Model Summary Sheet and Parameter Estimates Diagnosis chart prototypes were implemented as functional off-line prototypes. They both require additional development to be commercially viable, and of course can be improved further. To introduce an experimental evaluation of RE charts in Chapter 5, the next section summarizes the drawbacks and benefits of RE charts.

### 4.5.1. Remaining weaknesses of Recursive Estimates Charts

While the modifications made to conventional RE charts (Section 4.3.7) reduce some perceptual ambiguities, several weaknesses in the algorithm remain:

- RE charts are only straightforward to use with linear regression models that have positive coefficients. Negative model coefficients invert parameter chart behavior, which contradicts interpretation rules (Section 4.3.8). Negative coefficients can be avoided by redefining or pre-processing driver variables. For example, rather than a “Weekend” driver with a negative parameter coefficient, a model could instead include a “Workday” driver with a positive coefficient.
- RE charts will respond unpredictably to endogenous factors such as measurement error or autocorrelation (Section 4.3.4). Autocorrelation can be reduced by choosing a slower modeling frequency interval (e.g. daily rather than by-shift, or weekly rather than daily).
- Similarly, RE charts will not clearly indicate changes to parameters used in pre-processing driver data to make them more linearly associated with energy use. The most common example of such

an exogenous parameter is the temperature ‘break point’ used to convert outdoor temperature to heating degree-days (Aird, 1981).

- RE charts become less useful when used with driver variables that covary more. They may be less amenable than CUSUM to complex models with many non-independent parameters.

For these reasons, RE charts are not intended to replace CUSUM charts for reliably quantifying changes. Rather, they are intended to support strategy switches between detecting changes with CUSUM charts, initiating a diagnostic search based on RE charts, and continuing the search based on parameter information from the Model Summary Sheet. It is not clear whether RE charts will confuse or support practitioners. As discussed next, I argue they should at least do no harm.

#### 4.5.2. Extra value from same model

RE charts require no additional data or model cultivation activities, satisfying design criteria from Section 3.8.3. Interpreting RE charts, however, requires clear documentation of model-building choices and the meaning of energy driver variables, such as in a Model Summary Sheet (Figure 18). If energy models are fit to data that describes meaningful work domain structures such as Purpose-Related Functions production or heating (Section 3.3), RE charts can enable search of those structures (Section 3.8.2).



**Figure 26 - Diagnostic search starting points with CUSUM charts (left) are in terms of over/under consumption in time at a particular utility meter. RE charts expand diagnostic search in terms of energy model drivers often representative of more abstract system structure (right)**

As illustrated in Figure 26, rather than Comparative Analysis initiating diagnostic searches only from over/underconsumption, search can start with an association of over-underconsumption to particular business system processes (Section 3.8.1).

#### 4.5.3. Time-series Representation Aiding

In a companion article (Hilliard & Jamieson, 2014a) I argue that the modified RE charts meet most of the principles of Ecological Interface Design (Section 4.1.3) at least more so than CUSUM charts. First, RE charts more consistently map system properties to chart cues, particularly with the time-invariant weighting, informativeness mapping, and chart axis scaling modifications described in Section 4.3.7. Unlike canonical EID practice of deriving a priori the constraints of a particular system, RE charts empirically estimate more abstract ‘energy intensity’ properties through coincident correlation (though see Lau & Jamieson, 2006). The quality of the mappings thus depends on the quality of the Comparative Analysis model and its underlying data sources. The outcome will not be a perfectly diagnostic display, but hopefully an improvement in the usefulness for a given labor and equipment investment.

Second, RE charts should be interpretable as time-series signals, signs and symbols using worker skills, rules and knowledge (Rasmussen, 1986). The output of the symbol-processing RE strategy

is a set of time-series signals, which with practice in a particular system, should be interpretable using recognition and analogical ‘rules of thumb’ (Section 4.3.8). Knowledge-based behavior may be encouraged by each RE chart mapping onto a parameter in the symbolic Comparative Analysis energy model, explained by a companion Model Summary Sheet (Section 4.2) in terms of energy intensity of business processes.

Finally, the RE chart should support naturalistic strategy switch behaviors. Within Comparative Analysis strategies, the RE and CUSUM charts can be plotted simultaneously, allowing switching between CUSUM to robustly date and quantify change effects and RE to diagnose possible causes in terms of model parameters (Section 4.4). RE charts also expose more abstract system structure captured in the energy model (often Purpose Related Functions of Section 3.3). This can inform a Condition Survey strategy (Section 3.5.3) through triangulating energy waste to a process (PrF) whose associated equipment is measured at a particular energy meter (PFn) and therefore located in a certain area (PFo). The prototype display designs for RE charts and the Model Summary sheet arguably meet many of the principles of ecological interfaces, and were “explicitly designed on the basis of a detailed understanding of the work ecology” (Bennett & Flach, 2011, p. 137)

#### 4.5.4. Heuristic and Usability evaluation

The RE chart and Model Summary Sheet prototypes were evaluated heuristically and through focus group critiques (Hilliard et al., 2014) by professional energy analysts at the collaborating M&T software vendor. Since professional analysts can be expected to learn and apply any statistical method, they are not a pressing concern. Of more importance is whether novices’ M&T behavior will improve given access to RE charts, and whether non-statistically-inclined practitioners can over time learn effective ways to use them. Chapter 5 describes an M&T experiment to evaluate and contrast the benefits of CUSUM and RE charts to novice behavior in a synthetic task environment.

## Chapter 5

### Controlled evaluation of time-series change diagnosis support

#### 5.1 Motivation: M&T never studied as controlled environment

This chapter describes the first controlled experimental evaluation of human performance at energy Monitoring and Targeting (M&T) using Cumulative Sum of Differences (CUSUM) charts. The experiment also comparatively evaluated the prototype Recursive Estimates (RE)-based chart of Chapter 4.

##### 5.1.1. Goal 1: Study M&T performance with standard tools

The first motivation for this experiment was to evaluate the industry standard CUSUM-chart-based M&T structural change detection strategy. To evaluate CUSUM tools required developing measures to quantify effective performance, experimental procedures to enable participation by novice energy management practitioners, and scenarios with reasonable face validity to real work situations. Since strategy effectiveness depends on the work environment (Section 3.2.4), a goal of the experiment was to manipulate characteristics of the monitored system that had been observed to create ambiguous CUSUM charts (such as the scalloping, large, and overlapping changes of Section 2.5.6). Some of the comparisons that scenarios were crafted to explore were:

- Whether recurring effects of structural changes to energy ‘driver’ variables (such as a broken furnace that wastes fuel every winter-time) can be distinguished from new, unrelated changes?
- Whether larger changes are not just easier to detect, but also easier to diagnose?
- If overlapping changes are always harder to detect or diagnose?
- Whether changes associated with a driver/variable are harder to diagnose than changes associated with baseload/intercept?
- If M&V models used for quantifying large changes (See Figure 8) are interpreted similarly to models developed for M&T?

### 5.1.2. Goal 2: Assess new normative diagnosis strategy support

Second, this experiment was designed to evaluate the recursive estimates strategy-based work support tool developed in Chapter 4. Testing each interface<sup>8</sup> in an experimental condition can indicate whether a Recursive Estimates (RE) chart representation caused improved performance at distinguishing or diagnosing changes.

### 5.1.3. Experimental Hypotheses

Drawing from the work analysis (Chapter 3) and a pilot Experiment I (briefly described in Section 5.3), the following five hypotheses guided the design of this experiment:

- 1) RE charts will not have an effect on change detection.
- 2) RE charts will help participants commit fewer false alarms due to misidentifying recurrent effects of persistent changes (Rule 2 in Section 4.3.8).
- 3) RE charts will improve change diagnosis for non-seasonal driver variables, because CUSUM charts poorly distinguish driver and baseload changes (observed in Section 2.5.6).
- 4) Participants who draw linking lines between CUSUM charts and RE charts to diagnose will make fewer false alarms and more correct diagnoses (Section 4.3.9).
- 5) Change magnitude (CUSUM slope) will be a better predictor of detection and diagnosis than accumulated change evidence (CUSUM vertical displacement).

As discussed in Section 5.6.5 below, not all of these hypotheses could be conclusively supported or refuted by this experiment alone. However, they were used to guide choices in the design of the experimental method, and can guide future work.

## 5.2 Method: Synthetic M&T task and experimental design

Guided by the experiment goals and the above hypotheses, I developed a pencil-and-paper M&T task, including a synthetic system, scenarios, response syntax, and data analysis methods.

### 5.2.1. Stimulus development

The first step in developing the experiment was to simulate a natural system in which changes could be detected and diagnosed with M&T. The ‘cover story’ explained to participants was of

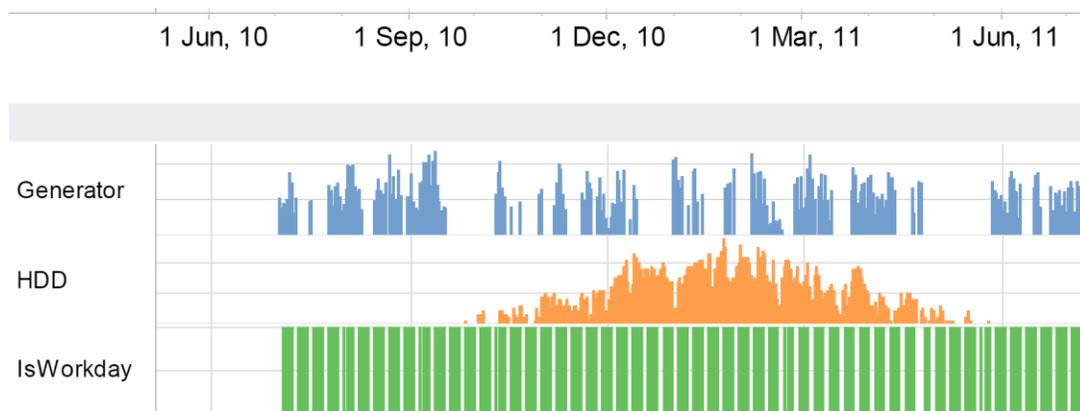
---

<sup>8</sup> Technically they are displays (lacking controls), I use “interface” for convention.

an office building consuming natural gas. Synthetic energy consumption data was developed from daily weather data recorded at the Toronto, ON international airport from 2009 – 2013. For the final Experiment II, synthetic energy consumption was calculated as the sum of:

- A. A fixed baseload gas consumption (intercept) of 3500 units (see Table 12)
- B. A fixed gas consumption of 5125 units on each workday (Monday-Friday except holidays)
- C. Gas consumption proportional to daily average temperature below 12°C, at 625 units per Heating Degree Day (HDD12)
- D. Gas consumption proportional at 615 units / h to a synthetic “hours operation” variable for a gas-electric generator (see below)
- Noise (described below)

Parameter values were chosen so that all four parameters would contribute equally on average to the simulated energy consumption (Table 12). The synthetic gas-electric generator operation data was generated by multiplying a uniform distribution (0..24h) time series with a binomial (on/off) time series. The uniform ‘operation’ time series was filtered by a recursive linear filter with coefficient of 0.5 to lengthen the timescale of values. The binomial (on/off) time series was filtered with a convolution linear filter with coefficients (5,1) to create visually distinctive continuous on/off periods. The product of these two is shown as the top graph in Figure 27.



**Figure 27 - One of the three years of data used for the Experiment II stimulus, including synthetic "Generator Hours" variable, Heating Degree Day (HDD) conversion of local temperature, and workday binary measure.**

**Table 12 – Experiment II Design of Stimulus Drivers and Parameters.**

<i>Driver</i>	<b>Min..Max Values</b>	<b>Mean Values</b>	<b>Parameter</b>	<b>Mean daily contribution</b>
<i>Generator</i>	0 .. 23.4 “h”	5.9	615	3610
<i>Temperature</i>	-14 .. 27 °C	9.2	N/A	N/A
<i>HDD12</i>	0 .. 26 °C-day	5.6	625	3502
<i>Workday</i>	0 .. 1	0.7	5125	3538
<i>Baseload</i>	1 .. 1	1	3500	3500
<i>Error (Weibull)</i>			1.5	293
<i>Total “Gas” Consumption</i>				14464

To obscure scenario features and increase task difficulty, the simulated energy consumption was obscured by three forms of variability:

- Gaussian ‘noise’ in all parameters (a mean of 10% for all terms)
- Gaussian error in the baseload (intercept) was slightly autocorrelated ( $\alpha = 0.2$ )
- A Weibull-distributed error term ( $k=1.5$ ) with mean of 8% of the baseload, representing occasional unmeasured energy waste

These noise terms were intended to challenge RE chart statistical assumptions (Section 4.3.4) and improve external validity. Ten percent variability was chosen through pilots and an Experiment I to create reasonably difficult-looking CUSUM charts. Noise terms were randomly generated between each scenario, but the same across experimental blocks and participants.

Next, following conventional model-fitting methods (Section 2.3.3), stable synthetic data dated July 2009 – July 2010 was used to train a regression model from which M&T energy performance scenario charts were generated (See Appendix D.1). This model typically fit with very high  $R^2 > .98$ . Depending on scenario, structural changes to parameters (Table 13) were introduced from late 2010 to late 2012. The next section describes the five experimental scenarios for the M&T task generated from this data synthesis process.

### 5.2.2. Scenario development

Five scenarios were developed for Experiment II, repeated in identical order in both normal ( $G_{1..5}$ ) and inverted ( $G_{6..10}$ ) groups of trials. Each of the ten resulting scenarios were composed of between one and four step changes to the underlying synthetic data parameters, labeled as types:

- A. Baseload
- B. Workday
- C. Heating
- D. Gas generator

The specific permutations of scenario change type (A..D), size (small/large), direction, and timing (Jun 2010-Jun-2013) used in this experiment were designed according to two principles:

- 1) The changes coordinate to create emergent phenomena I observed being difficult to diagnose (Section 2.6.3):
  - a Changes to intermittent (workday, generator) drivers confusable with change in baseload
  - b Large change in baseload obscuring subsequent smaller changes to parameters
  - c Multiple changes occurring simultaneously
  - d Persistent changes overlapping each other
- 2) The changes were balanced in size and quantity. Each set of 5 scenarios had
  - a 3 changes to each of the four parameters, of which
    - i. 1 change leads, and 2 occur later in the scenario
    - ii. 2 are small, and 1 is large.
  - b Accumulated change evidence<sup>9</sup> (effects of change size over the duration before the next change) ranging between 2.4 and 22.5 average days' consumption.

The set of changes used in Experiment II scenario blocks  $G_{1..5}$  and  $G_{6..10}$  are listed in Table 13 and in Appendix F.1. The resulting scenarios generated for Block  $G_{1..5}$  are illustrated in Appendix F.3.

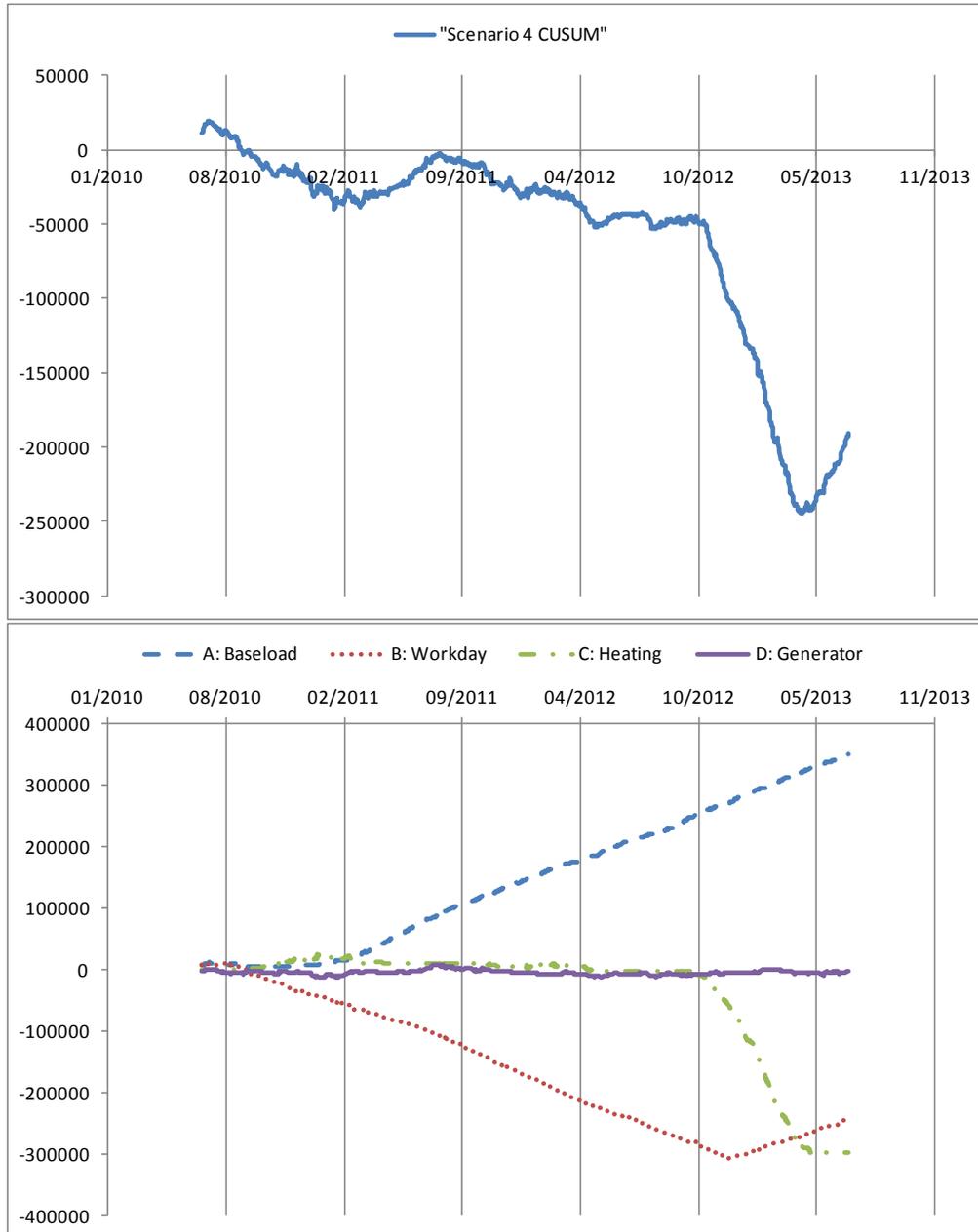
---

<sup>9</sup> Accumulated evidence was calculated as the product of parameter change (e.g. 10%), and the accumulated driver variable values over the uninterrupted time before the next change or the end of the scenario, normalized by the average gas consumption. It is expressed in terms of average number of days consumption saved/overconsumed.

**Table 13 – Properties (at left) of all changes in Experiment II Scenarios (at right). Leading changes appear first (labeled with alphanumeric \_A) in the scenario. Large changes are double small changes. Type/Cause refers to which parameter changed (A..D). Accumulated evidence is a product of change size, driver variables, and intervening time between changes. Change 2A/7A was not scored as it served only to obscure subsequent changes (2B/7B and 2C/7C).**

	Leading change (first in scenario)	Large size change (20% vs. 10%)	Type/Cause of change	Accumulated Evidence (average days' consumption)	TrueChange (G <sub>1..5</sub> )	TrueChange (G <sub>6..10</sub> )	
<i>Properties of Changes (at right)</i>	TRUE	TRUE	A	2.4	3A	8A	
		FALSE	B	4.9	4A	9A	
			C	22.5	5A	10A	
			D	15.0	1A	6A	
	FALSE	TRUE	B	9.9	4D	9D	
			C	4.9	4C	9C	
			D	18.9	3B	8B	
		FALSE	FALSE	A	9.9 9.5	1B 4B	6B 9B
				B	9.9	3C	8C
				C	10	2B	7B
				D	5.5	2C	7C

An example showing how the effects of four persistent changes 4A, 4B 4C and 4D combined to create a Scenario 4 CUSUM chart is shown in Figure 28. Of course, a dis-aggregated chart of each change is not available in practice, and was not provided to experimental participants.



**Figure 28 – Accumulated Evidence of four persistent changes, 4A (-10% workday), 4B (+10% baseload), 4C (-10% heating), and 4D (+20%, workday) superimpose (bottom) and sum to create the Experiment II Scenario 4 CUSUM chart (top). “Generator” performance did not change in this scenario.**

### 5.2.3. Experimental design

The experiment was structured as a 2x2 within-subjects design with two levels of Interface and Scenario that participants experienced in different orders. A within-subjects design was chosen since representative experimental participants were scarce (Oehlert, 2000). Participants were assigned by block randomization to one of 4 trial orders (Table 14). All participants experienced both levels of two fixed factors  $T_C$  or  $T_{C+R}$  (interface design treatment) and  $G_{1..5}$  or  $G_{6..10}$  (group

of scenarios). Participant orders counterbalanced learning effects of **T**. Scenario groups **G** were composed of  $i=5$  ordered trials of near-identical scenarios  $S_i$ . Scenarios in  $G_{6..10}$  were inverted and generated with different random variation to obscure their similarity to  $G_{1..5}$  and reduce learning effects.

**Table 14 - Experimental Design of Blocks and associated Treatment and Scenario group Order**

<i>Participant</i>		<b>Order</b>	<b>First i trials</b>	<b>Second i trials</b>
$P_{1+4n}$	Block Random Assigned to	A	$T_C, G_{1..5}$	$T_{C+R}, G_{6..10}$
$P_{2+4n}$		B	$T_{C+R}, G_{6..10}$	$T_C, G_{1..5}$
$P_{3+4n}$		C	$T_C, G_{6..10}$	$T_{C+R}, G_{1..5}$
$P_{4+4n} \dots$		D	$T_{C+R}, G_{1..5}$	$T_C, G_{6..10}$

#### 5.2.4. Response forms and equipment

The scenarios were presented to participants as two 11x17" paper booklets of 5 charts each with an associated response table (attached in Appendix F). Trials in the  $T_C$  condition contained time-series charts of driver data, energy consumption, predicted energy consumption, Control Chart, and CUSUM charts. Trials in the  $T_{C+R}$  condition contained the same  $T_C$  charts, plus four Recursive Estimates charts (formatted as described in Section 4.3.7), one for each of the model drivers (which were described in Section 5.2.1). Participants were given:

- A red pen (for answer-marking)
- A pencil and eraser (for drawing all other marks, references, and perceptual aids)
- A 30cm clear ruler (for perceptual aid tactics described in Section 4.3.9)
- Two strips of construction paper 11" by 2", to be used as masks (for focused attention tactics described in Section 4.3.9).

#### 5.2.5. Participants

Final-year students from energy building systems programs were recruited from Humber College and Seneca College in the Toronto area. Ethics approval was obtained from the University of Toronto and both colleges, and students were contacted in a research ethics board-approved manner.

Compensation for each participant was CAD \$15 / hour + \$10 for completion. All participants completed the task and most students required less than 2 hours. Median compensation was \$40. Recruitment was anticipated at 24 participants. Experiment I attracted only 19 but Experiment II was better scheduled and 33 participated in total (Table 14).

### 5.2.6. Administration method

Experiments were conducted on-site at participating colleges, in a classroom setting. Each session began with an introductory presentation by the experimenter covering (in Experiment II):

- 1) Overview of RE charts – how they're generated, what they mean
- 2) Two suggested modes of interpreting RE charts (described in Section 4.3.8)
- 3) Suggested tactical rules for responding (using a ruler to determine and link change points, described in Section 4.3.9)
- 4) Three practice trials, which each participant completed individually
- 5) Group review of true answers for each practice trial
- 6) An overview of the study purpose, expected time commitment, and compensation
- 7) Opportunity to discuss and decide whether to complete the Informed Consent Form
- 8) Review of written instructions, and answering clarifying questions. Participant instructions are attached as Appendix F.2.

Subsequently, participants were seated individually (as for a quiz), and the first trial booklet distributed sequentially by seating order from the front of the class. Randomization occurred through the block randomization of the trial booklets to ordered interface and scenario groups (see Table 14). Participants were asked to complete each booklet using a supplied pencil, ruler, and red pen. When complete, they were asked to call the experimenter (or assistant) to verify the completeness and coherence of their charts and response table responses. Start time of each group, and completion times for individual booklets were recorded by the experimenter and assistant using a synchronized clock.

### 5.2.7. Data Interpretation and Entry

Participants' handwritten response tables and charts were input into an Excel spreadsheet by hand. An example is shown in Figure 29. Response tables were interpreted with the following

rules, intended to capture the intent of participants, even if they did not respond exactly according to task instructions:

- If a cell was left blank, record N/A (not available)
- If a range of dates was indicated, record the month at the midpoint of the range
- If two diagnosis attempts marked on response sheet,
  - ◆ Record the one that's not "do not know / cannot tell"
  - ◆ If both are causes, record "do not know / cannot tell "

Chart marks in red ink were measured with a ruler to the nearest half millimeter from the Y-axis. Interpretation rules were:

- If no "X" marked, measure pencil-sketched vertical line. Otherwise, use midpoint of marked response window.
- If no response window marked, use width of X (or a half-mm).
- If markings split between two charts, measure X and response window separately. Record chart number where response window marked.
- If redundant marks on two charts, measure the one with the greater span.
- If two response window interval line markings, measure outside (greater) one.
- If vertical lines slanted / skewed, measure to intersection with horizontal line

Data entry and consistency of the interpretation rules were validated for Experiment II by a 10% sample re-processed by independent investigators. Krippendorff's Alpha for all measures was over 0.95, and of 20 mismatches, 17 were typos by the independent investigators and 3 ambiguous responses by participants (see Section 5.5.5.1).

Participant	209												c
Scenario	10 of 10												
Suspected Change Times	(Mark as many changes as you judge important and label each with a different letter)												
Labeled With	A	B	C	D	E	F	G	H	I	J	K	L	M
Approximate Month / Year (Of the "X" you marked)	Feb/10	Nov/11	Nov/12										
How confident are you that this is a new change event? (0 = not at all, 10 = completely)	8	5	8										
Effect on Energy performance	(Mark either - or +)												
Did the change Save (-) or Increase (+) consumption?	+	+	+										
Suspected Cause	(Mark the one that best explains each change)												
a) Not certain / Can't tell													
b) Change in consumption rate of natural gas generator	X												
c) Change in furnace, HVAC, or building insulation envelope													
d) Change in workday - related processes or equipment													
e) Change in always-on or everyday equipment		X	X										

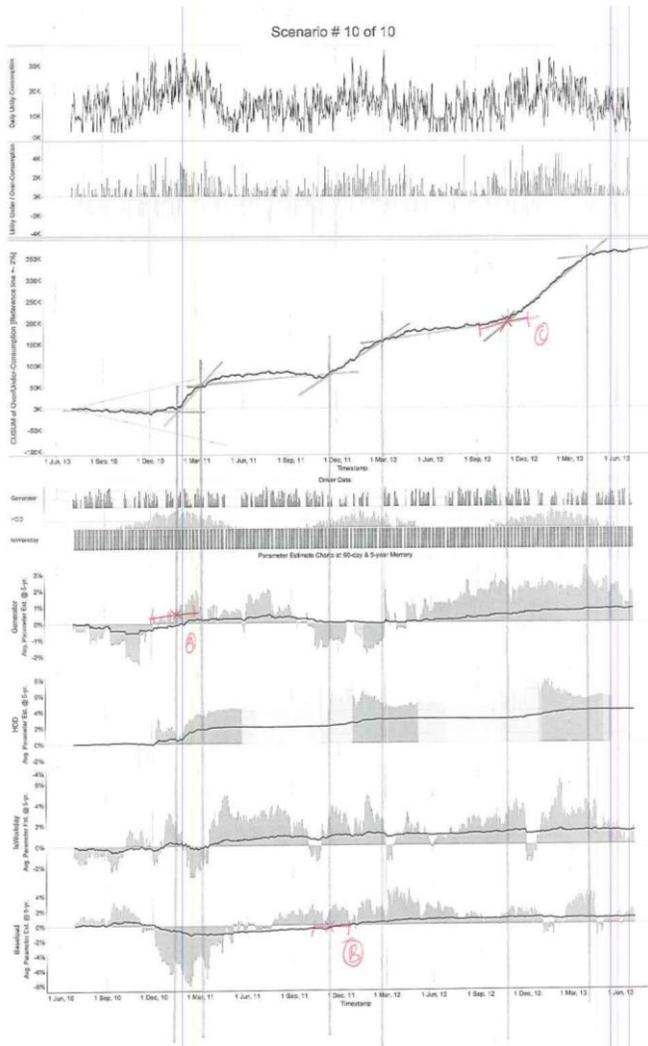


Figure 29 - Example response from Experiment II, Participant 209, Order C, Trial 10, Scenario 5, G<sub>1.5</sub>.

### 5.2.8. Data Scoring

Participants' responses were scored against the 'true changes' template for each scenario according to an R script. As with the response interpretation rules, the goal of the scoring rules were to reasonably interpret participants' intent. Two scoring rules were developed: a conservative "InBox", and a more liberal "Likely" rule. The algorithm for each is outlined in Table 15.

**Table 15 - Response  $\leftrightarrow$  Change scoring rules, recursively applied for each trial until no "best" pairings left**

<i>Criteria</i>	<b>'InBox' rule</b>	<b>'Likely' rule</b>
<i>What is the scoring rule intended to do?</i>	Strictly interpret according to task instructions	Forgive slightly early, fairly late and over-confident responses
<i>Within each scenario, how 'good' are the eligible response-change pairings? (sort by first to last)</i>	Response confidence window overlaps change Response "X" is after change Response "X" is closest to change Response direction matches change Response diagnosis matches change	
<i>Does the 'best' response-change pairing count as a 'Hit'?</i>	Response not already a 'hit' Change not already 'hit' by a response AND	
	Response confidence window overlaps change  OR Response "X" is within +/- 7 days of change	OR Response "X" is within +/- 14 days of change OR Marked after AND as same direction AND no intervening false alarms.
<i>Which response-change pairings should be dropped from the eligible list?</i>	All pairs involving the change from the "best" pair, AND: If "best" a hit, all other pairings of that response If "best" not a hit, all equally or more distant pairings.	

Which scoring rule is best for experimental data analysis? I scored the analysis described here using the "Likely" rule, since it more closely matches the task constraints, and should reduce effects of participant over-confidence (narrow response windows) on performance. In the M&T

task, detecting changes late is still worthwhile, as long as participants are not falsely responding frequently. In practice, appropriate confidence is more important to avoid false alarms. Therefore, the experimental task was designed to reward appropriately confident behaviour of marking generously wide date ranges in uncertain cases. However, participants in pilot Experiment I consistently marked short time ranges, missing the true change date. Since participants' confidence (mis-)calibration is not the primary focus of this experiment, a scoring rule that forgives over-confident narrow responses is appropriate.

### 5.2.9. Data Categorization and Measures

Response data was interpreted using the above rules and categorized in detection and diagnosis at six levels of aggregation. The categories used to describe response data were inspired by the signal detection theory (SDT) framework (Macmillan, 1991), and are described in Figure 30 and Table 16. Every response (R) marked by participants is either with (AD) or without (ND) an attempted diagnosis. Responses that are not scored as matching a scenario change (Section 5.2.8) are false alarms (FA). Responses that do match a change are categorized as Hits (H), either with (HD) or without (HN) a diagnosis. Some of the hits with diagnoses will be right (RD).

Signal detection theory serves as a conceptual framework, but does not completely apply to this experimental design. While the definitions of Hits, False Alarms, and Misses are compatible (see a decision flow chart in Figure 30), Correct Rejections are not defined for the continuous multi-target stimuli used in this experiment for two reasons. First, in principle Correct Rejections can be infinite, since participants did not know the maximum number of true changes present (the response form had space for nine). This might be overcome by analyzing time segments of each scenario independently, but persistent overlapping changes meant that each segment would not be perceptually independent of preceding segments, a violation of SDT assumptions. Undefined correct rejections unfortunately means that measures of sensitivity  $d'$  and bias  $\beta$  cannot be computed. However, as outlined in Table 17, ratios of defined SDT-inspired parameters can provide an alternative.



**Table 16 - Taxonomy for M&T experiment data, described as a cross-tabulation of responses matched to true changes. Detection (H, FA) and Diagnosis (AD, ND) categories overlap. Misses are not shown in this table, and Correct Rejections do not apply to this experiment. See Table 22 for experimental data in this format.**

<b><i>R: Responses can be categorized as combinations of:</i></b>	<b>H: Hit = HD + HN</b>	<b>FA: False Alarm = FD + FN</b>
<i>AD: Attempted Diagnosis = HD + FD</i>	HD: Hit with Diagnosis Attempt <div style="border: 1px solid black; padding: 5px; display: inline-block;">RD: Right Diagnosis</div>	FD: False alarm with Diagnosis attempt
<i>ND: No Diagnosis attempt = HN + FN</i>	HN: Hit with No diagnosis attempt	FN: False alarm, No diagnosis

**Table 17 - Data Analysis ratios of categories shown in Table 16 and Figure 30. These ratios can 1) correct for and/or 2) quantify participants' propensity to a) respond (detection) and/or b) identify a cause (diagnosis). Ratios can also describe the proportions of attempted diagnoses or hits broken down by change type/cause.**

<b><i>Data Analysis Ratio</i></b>	<b>Abbreviation</b>	<b>Formula</b>	<b>Propensity / Bias</b>	<b>Detection Performance</b>	<b>Diagnosis Performance</b>
<i>Hit Rate</i>	Hr	H/TC	X	X	
<i>Right Diagnosis Rate<sup>10</sup></i>	RDr	RD/HD			X
<i>Correct Response Rate</i>	CRr	RD/R	X	X	X
<i>False Alarm Rate</i>	FAr	FA/R	X	X	
<i>Attempted Diagnosis Rate</i>	ADr	AD/R	X		X
<i>Hit and Diagnosis Rate</i>	HDr	HD/H	X		X

<sup>10</sup> Right Diagnosis rate is the only ratio that does not inflate with propensity to respond or attempt diagnosis.

Counting responses (R) or true changes (TC) according to these categories is not sufficient. Ratios of response categories are better performance measures and can isolate task steps. Ratios used in this data analysis are outlined in Table 17. However, these ratios are not ideal performance measures. Participants' propensity to respond (their response bias) will inflate most of these ratios, such as both measures of detection performance, Hit rate (Hr) and False Alarm Rate (FAR). Unfortunately bias-independent measures of detection (such as  $d'$ ) are not defined for the perceptually overlapping stimuli used in this experiment (Macmillan, 1991). The least bias-inflated measure is diagnosis performance indicated by Right Diagnosis Rate (RDr).

**Table 18 - Data Aggregation Levels used in analysis of Experiment II data. Data becomes more aggregated from left to right (by-participant) and from top to bottom (by-experimental-stimulus).**

		Increasing Aggregation by Response measures →					
Increasing Aggregation by Change measures →	All Responses (N=1329)	<	By Participant and Trial (n=330)	<	By Participant and Condition (n=66)	<	By Participant (n=33)
	^		^				
	By Each True-Change (n=26)		By TrueScenario (n=10)	<	By Scenario (n=5)		
	^						
	By each Change (n=13)						
	^						
	Each Change-Type (n=4)						

Each data category or analysis ratio may be aggregated, depending on the questions to be answered, from individual responses up to summary sets (e.g. by-participant), as outlined in Table 18. Total responses can be aggregated two main ways: by participant performance, i.e. this participant detected 50% of changes (Hr) they were presented; or by change totals, i.e. this change was detected 50% of the times it was presented (pH) to participants.

### 5.2.10. Questionnaire

The final experimental measure was a one-page questionnaire, developed to elicit Experiment II participants' opinions on the CUSUM and Recursive Estimates charts. It consisted of 10 five-point Likert scales (Strongly Disagree to Strongly Agree) and a free-form comment box (see Appendix F.3 ). The Likert scales contrasted participants' opinions on whether CUSUM and RE charts were:

- 1) easy to understand,
- 2) clear in showing when changes happened,
- 3) clear in showing what type of changes happened,
- 4) informative, and
- 5) confusing.

The questionnaire was administered to each participant after completion of Experiment II.

### 5.2.11. Statistical Methods

Analyses described below were performed with simple non-parametric tests, but also with mixed-effects generalized linear regression models. Mixed models were chosen since they distinguish random effects of participants and schools, and correctly account for repeated measures from the same participants (Oehlert, 2000). Generalized models were used since they are more robust to unbalanced data with missing values since participants

- were scarce
- might not be recruited in perfect balance each school and counterbalancing order
- could identify different numbers of changes and
- were expected to make some errors in responding.

Generalized logistic modeling was used since it could interpret un-aggregated response data to model the likelihood of binary outcomes (e.g. Hit or Not).

## 5.3 Results, Pilot Experiment I

The first attempt at an experimental evaluation failed to recruit sufficient subjects and had design limitations (discussed below in Section 5.4). Experimental power was low, and few significant

results were found. Results are summarized below in less detail than those of Experiment II (Section 5.5)

### 5.3.1. Participation

Experiment I recruited eighteen participants. Data quality for two participants was questionable. Participant #106 repeatedly marked four changes per trial, one per RE chart, and was excluded for misunderstanding the task. Participant #111 did not complete the “change direction” field in the response table, and attempted no diagnoses in the RE chart condition. Participant #111’s data was retained, as zero diagnoses was not an outlier.

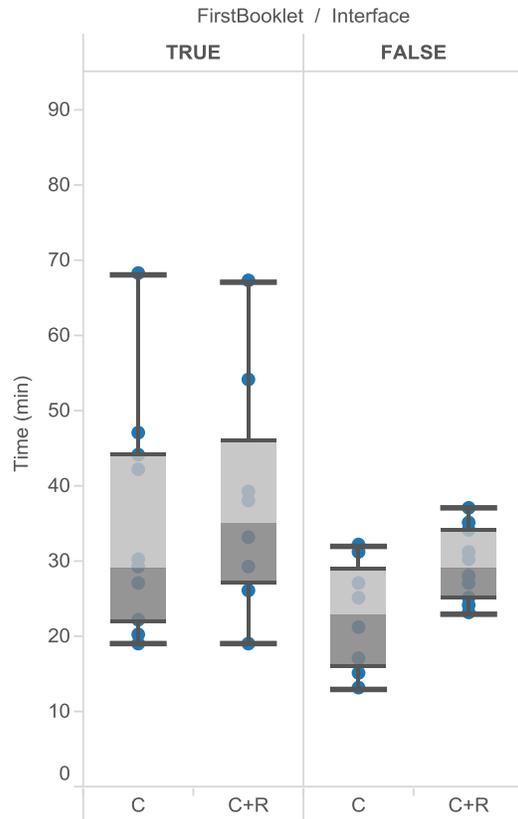
**Table 19 - Experiment 1 Participants by experimental block Order and School. Asterisk\* indicates one participant (106) removed due to misuse of RE charts.**

<i>Order</i>	<b>School HC</b>	<b>School SC</b>	<b>Total by Order</b>
<i>A</i>	2	3	5
<i>B</i>	1	2*	3*
<i>C</i>	2	3	5
<i>D</i>	3	2	5
<i>Total by School</i>	8	10	18

As a result of excluding participant #106, most experimental block orders had 5 participants, but Order B had only three participants (Table 19). Order B’s sample size of three makes it difficult to assess interaction effects of a participant first receiving scenario booklet  $G_{6..10}$  with the C+R interface.

### 5.3.2. Response time and types

Participants took on average 31.3 minutes ( $SD = 12.6$ ) to complete each booklet of 5 scenarios (see Figure 31). A Wilcoxon matched-pairs signed rank test indicated a significant difference between the first ( $n = 18$ ) and second ( $n = 18$ ) booklet a participant completed,  $W = 27$ ,  $p = .01$ ,  $r = .38$ . Participants took a median of 8.5 minutes longer, 95%  $CI [16.5, 2.0]$  to complete their first booklet. Participants did not take significantly longer  $p > .05$  to complete booklets with the C+R interface (See Appendix C.3.1).

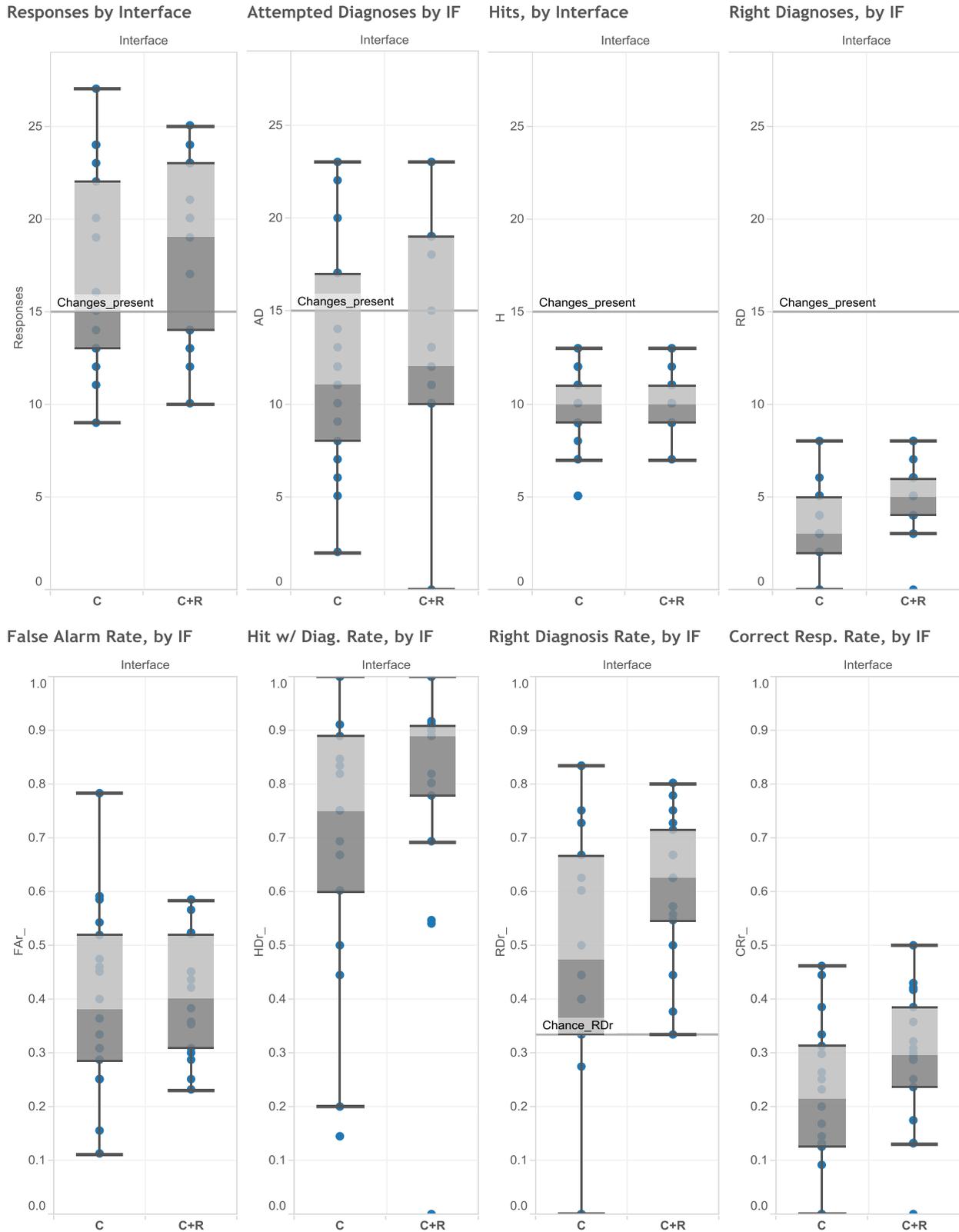


**Figure 31 - Time taken in Experiment I to complete each booklet of 5 scenarios ( $N=32$ ), according to first booklet (outer sort) and whether the scenarios were displayed with CUSUM-only or CUSUM+RE charts (inner sort).**

Participants marked 634 responses in total, and marked quite short response windows ( $Mdn = 48$ ,  $M = 58$  days), corresponding to pen marks 0.9cm wide. Participants responded most often on the CUSUM chart in both experimental conditions, but used RE charts almost as often as CUSUM charts in the C+R condition (See Appendix C.3.2).

### 5.3.3. By-Participant performance and Interface effects

According to the “likely” change scoring rule (see Section 5.2.8), Participants on average hit 10 ( $SD = 2.0$ ) of 15 changes (67%) in each five-scenario booklet. However, at the same time they committed 7.6 false alarms ( $SD = 4.3$ ). Participants responded significantly more often ( $M = 17.6$ ) than there were true changes (15) in each scenario booklet,  $W = 448$ ,  $p = .01$ ,  $r = .39$ . The distribution of responses is plotted in Figure 32.



**Figure 32 – Experiment I Detection and Diagnosis summary, according to “Likely” scoring rule, aggregated by CUSUM-only or CUSUM+RE experimental condition. Top row are counts of R, AD, H, RD. Bottom row are rate-normalized according to Table 17.**

### 5.3.3.1. Detection Interface Effects

Interface condition had no significant effect on hit or false alarm rates. Participants responded comparably often in the C condition ( $M = 17$ ,  $SD = 5.2$ ) as the C+R condition ( $M = 18.2$ ,  $SD = 4.9$ ), and hit a comparable number of changes ( $M = 9.7$  vs.  $M = 10.4$ ). Wilcoxon matched-pairs tests and mixed effect logit models did not find these differences significant (See Appendix C.4.1).

### 5.3.3.2. Diagnosis Interface Effects

Participants attempted diagnosis at comparable rates with both the C interface (72%) and in the C+R condition (74%). Diagnosis success rates for detected changes were comparable in the C condition (52%) and C+R condition (60%). Both were significantly better than chance (33% in Experiment I) according to a Wilcoxon matched-pairs signed rank test,  $W = 143$ ,  $p = .01$ ,  $r = .53$ . However interface conditions did not significantly differ according to mixed-effects logit models of likelihood of both diagnosis attempts (ADr),  $p > .47$  and right diagnoses (RDr),  $p > .20$  (See Appendix C.4.2).

## 5.3.4. Stimulus Effects

Stimulus properties were difficult to disentangle in Experiment I. Unlike the stimulus developed for Experiment II (Section 5.2.2), the three contributors to changes in Experiment I were given different magnitudes (See Appendix C.1). Baseload accounted for 70% of the Experiment I synthetic energy consumption stimulus, Heating 25%, and the third driver (precipitation) only 5%. This size difference was partially compensated by changing scenario change size (e.g. a small change to baseload was 5%, but 11% to heating and 15% to precipitation). Nevertheless, changes to baseload would be expected to be more detectable (and distinct) than changes to precipitation, just by CUSUM slope magnitude alone.

### 5.3.4.1. By-Change Properties Detection effects

The influence of change properties on detection were modeled with mixed-effects logistic regression. Through reverse fitting (See Appendix C.5.1), the most likely model described detection ( $p < .02$ ) in terms of changes that were: Caused by Heating factor  $OR: 3.6$ , 95%  $CI$  [2.1, 6.3], Large sized  $OR = 4.9$ , 95%  $CI$  [3.0, 8.1], and Leading  $OR = 0.56$ , 95%  $CI$  [0.34, 0.90].

The ‘accumulated evidence’ property was borderline significant, but difficult to isolate as it was highly correlated with Change type (due to the Experiment I stimulus design).

#### 5.3.4.2. By-Change Properties Diagnosis effects

The most likely mixed-effects logistic regression model describing likelihood of a change being rightly diagnosed (RDr) was a significant interaction  $p < .001$  between Interface type and Change cause. Statistical output is listed in Appendix C.5.2. Contrasts found weak evidence  $p > .04$  that the C+R Interface may have been associated with worse diagnosis of change type A (baseload) and better diagnosis of change type C (precipitation, contributing just 5% to consumption).

However, results were inconclusive.

### 5.4 Pilot Experiment I Discussion

Low participation rates contributed to Experiment I’s inconclusive results. Right diagnosis rates were encouraging, over 50%, but it was not clear whether this was due to participant skill, differing sizes of the three change types, or there being only three diagnosis options. From examining and analyzing results, I identified some improvements to the experimental procedure to better separate the effects of change size and change type / chart shapes. The experimental methods discussed in Section 5.2 incorporate these changes drawn from Experiment I:

- A fourth energy driver was introduced (Weekday), and the intermittent energy driver (Precipitation) replaced with a less peaky synthetic variable (Generator) (Section 5.2.2)
- Scenarios were formulated to have equal contributions of each energy driver, and more systematically varied change properties (see Table 13), to distinguish change size from the resulting characteristic chart shape
- Participant training was expanded with a third practice trial, and more progressive practice difficulty
- Instructions were changed to ask participants to mark an “X and response window” instead of a response window alone. This increased redundancy and helped interpret intent (Section 5.2.7)
- Stimulus formatting was adjusted to increase RE short-timescale (grey) chart salience.
- Participants were provided with black cardboard “viewport” strips to (optionally) use to mask off RE charts to focus on times of interest

## 5.5 Results, Experiment II

The final experiment was conducted without incident, according to the methods described in Section 5.2. This section first introduces the measures used to describe the data, then summarizes the participants recruited. Next, participants' performance is discussed in terms of detection and diagnosis. Change detectability and diagnosibility are also analyzed in terms of cause type. Finally, influences of data entry, scoring rules, learning effects, guideline-sketching, and diagnosis attempts are explored. The section concludes with a summary of participant feedback.

### 5.5.1. Participation

Thirty-three students participated, 22 from Humber College, and 11 from Seneca College (Table 20). Students were randomly block assigned to orders, with 8 in each order except 9 in A.

**Table 20 - Experiment 2 Participants by experimental block Order and School**

<i>Order</i>	<b>School HC</b>	<b>School SC</b>	<b>Total by Order</b>
<i>A</i>	5	4	9
<i>B</i>	5	3	8
<i>C</i>	6	2	8
<i>D</i>	6	2	8
<i>Total by School</i>	22	11	33

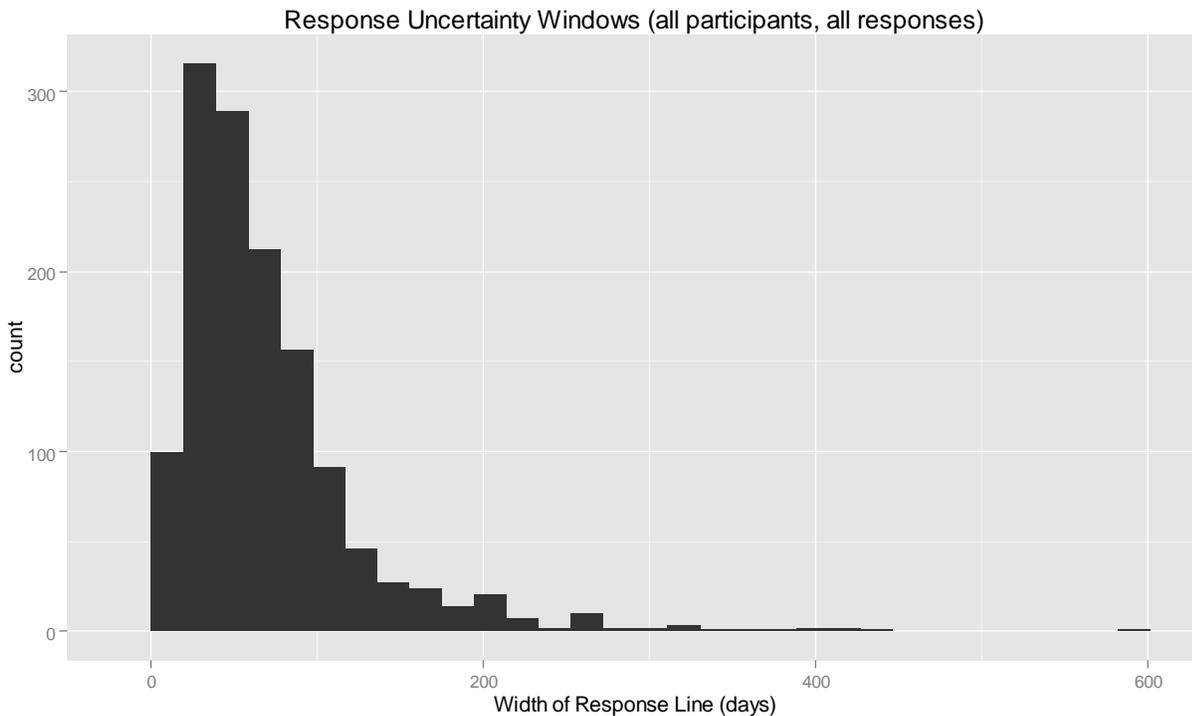
Most participants responded according to instructions, and ambiguities were resolved according to the rules in Section 5.2.7. Two participants (202 and 221) seemed to misinterpret the instructions and marked the 'duration of influence' of a change rather than the uncertainty in the onset date. Despite this, all participant data was retained to be representative of student practitioners-in-training as a whole.

### 5.5.2. Response types and time

During the experiment, we observed that participants appeared to make a serious effort to complete the experimental trials, and consulted the Recursive Estimates charts in the C+R condition. The number of participants that used the provided rulers and black cardboard chart masks was not observed.

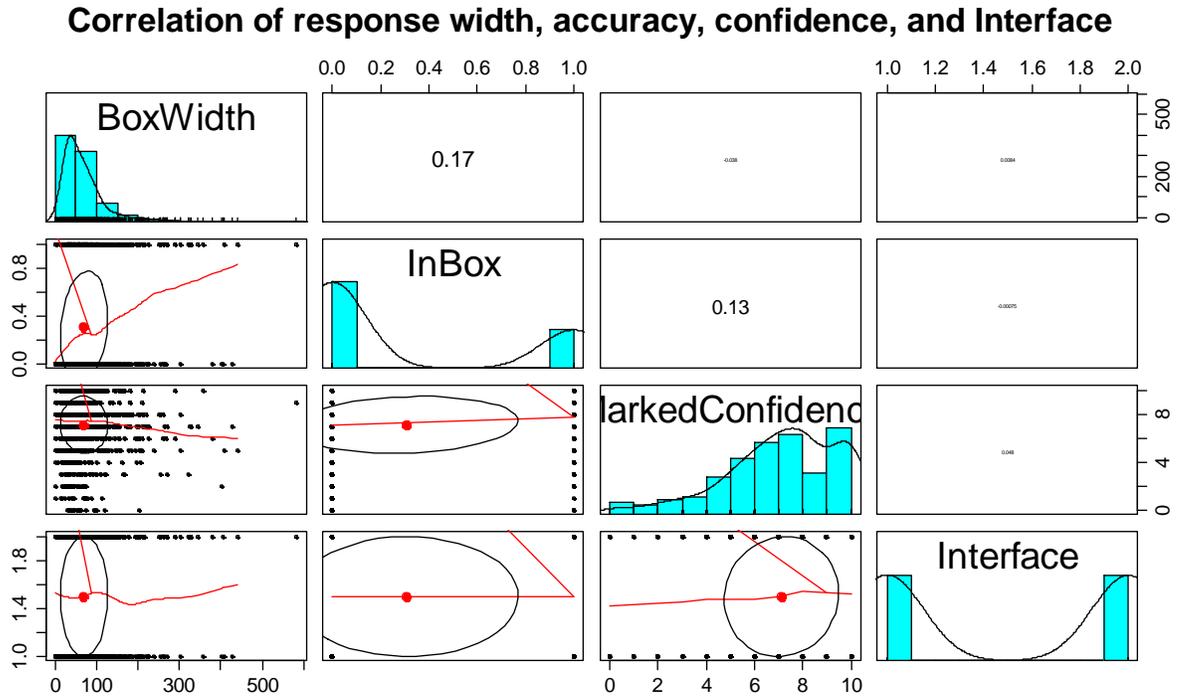
### 5.5.2.1. Response Windows

Thirty-three participants marked 1329 responses in total (Table 18). Contrary to instructions to mark responses “long enough to cover the date range when the change could have happened” participants marked quite short response windows ( $Mdn = 55$ ,  $M = 70$ ), corresponding to marks less than 1.5cm wide (Figure 33). Many changes were marked with a response window of less than 20 days / 4mm.



**Figure 33 - Width of response windows, for all participants/scenarios/conditions. A response width of 100 days corresponds to a 1.9cm line drawn on a scenario chart.**

Response window width was not significantly correlated with indicated confidence  $r(1316) = -.04$ ,  $p = .17$ , hit likelihood, or interface used (Figure 34). This is consistent with the instructions to draw the window as long as needed to be confident.



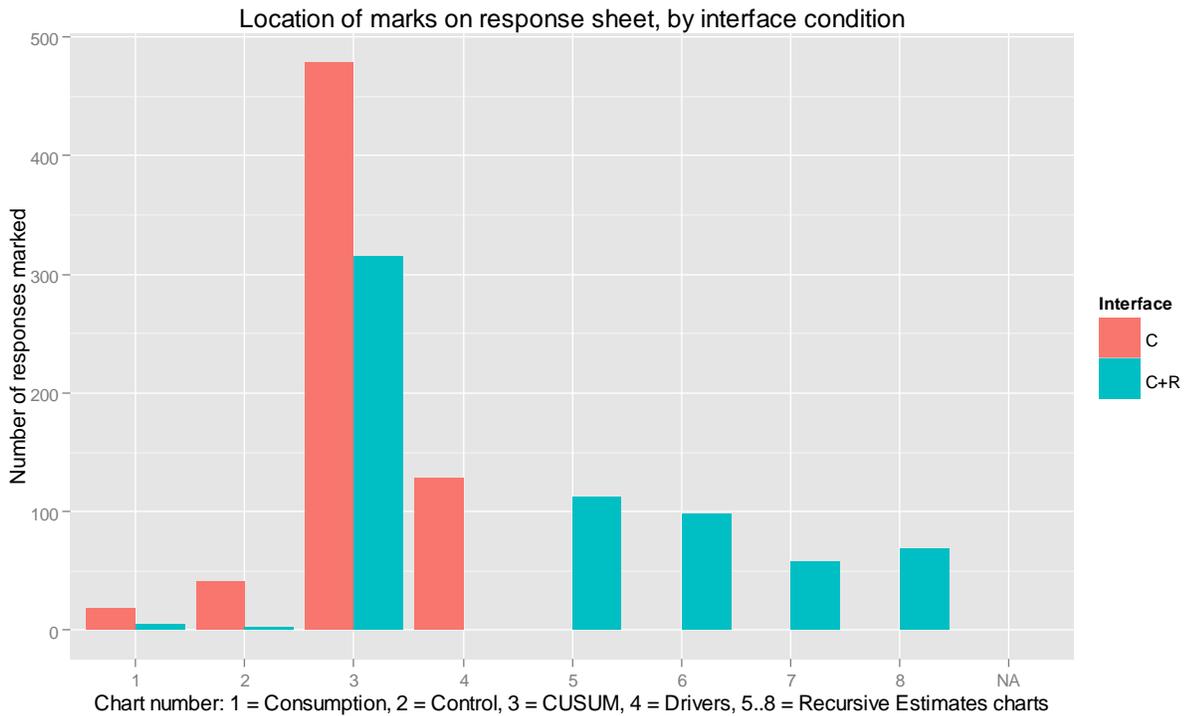
**Figure 34 – Pearson correlations ( $N=1329$ ) between response window width (BoxWidth), whether the response overlapped a change (InBox), the indicated confidence in the response, and the Interface condition.**

#### 5.5.2.2. Response locations

Participants were instructed to respond on the “chart you think most clearly shows the change”. In the C condition, they mostly (72%) relied on the CUSUM charts (see Table 21). In the C+R condition, they indicated changes on the four RE charts about as often (51%) as CUSUM charts, and almost never (1.3) responded on driver variable, control, or consumption charts. Across both conditions, the CUSUM chart remained the most popular (see Figure 35).

**Table 21 - Response locations for all participants ( $n = 33$ ), all scenarios ( $s = 5$ ) in each experimental condition (CUSUM only, or CUSUM+RE). Charts are numbered as they appear on the response forms, from top to bottom.**

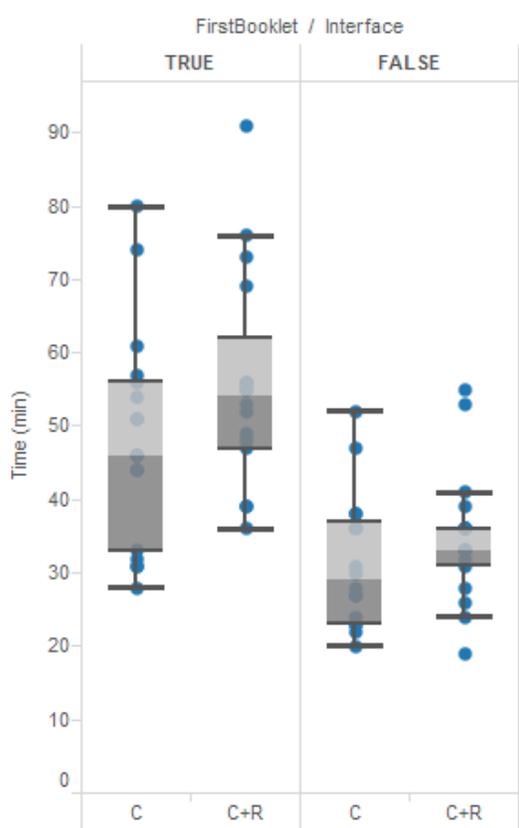
Interface Condition	Response Sheet Chart Number								<b>Chart numbers:</b> 1: Consumption and Modeled 2: Control 3: CUSUM 4: Driver variables 5: RE - Generator 6: RE - Heating 7: RE - Weekday 8: RE - Baseload
	1	2	3	4	5	6	7	8	
CUSUM	18	41	479	128					
CUSUM+RE	5	3	315	1	113	98	58	69	
Total	23	44	794	129	113	98	58	69	



**Figure 35 - Location of response marks on response sheet, by interface condition (CUSUM or CUSUM + RE)**

### 5.5.2.3. Time taken

Participants took on average 42.5 minutes ( $SD = 15.6$ ) to complete each booklet of 5 scenarios. A Wilcoxon matched-pairs signed rank test indicated a significant difference between first ( $n = 33$ ) and second ( $n = 33$ ) booklets,  $W = 27$ ,  $p < .001$ ,  $r = .54$ . Participants took a median of 17.5 minutes longer, 95%  $CI [24.5, 11.0]$  to complete their first booklet.



**Figure 36 - Time taken in Experiment II to complete each booklet of 5 scenarios ( $N=66$ ), according to first booklet (outer sort) and whether the scenarios were displayed with CUSUM-only or CUSUM+RE charts (inner sort).**

Participants took a median of 5.0 minutes longer, 95%  $CI$  [10.0, -1.5] to complete booklets in the C+R condition ( $n = 33$ ), not significant according to a Wilcoxon matched-pairs signed rank test,  $W = 220$ ,  $p = .28$ ,  $r = .07$ . Applying a more powerful mixed-effects generalized Poisson regression, adjusted for overdispersion, the effects of FirstBooklet ( $z = 3.14$ ,  $p = .002$ ) and Interface were significant ( $z = 1.97$ ,  $p = .049$ ). However, a likelihood ratio test found that Interface only improved the model at  $\chi^2(1, N = 66) = 3.77$ ,  $p = .052$ ,  $\phi = .24$ . See Appendix D.3.1 for full statistical outputs.

### 5.5.3. By-Participant performance and Interface effects

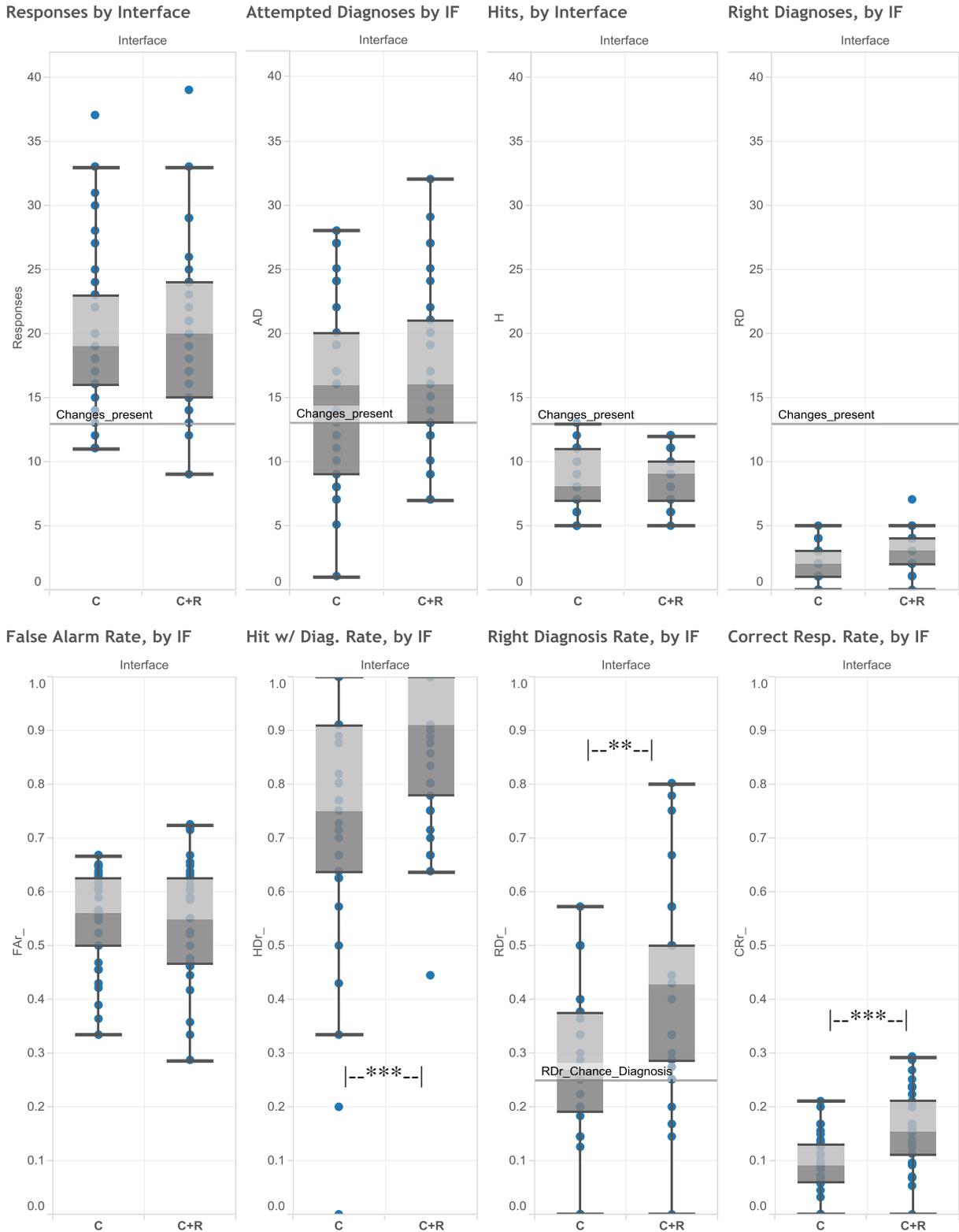
Detection performance was described in terms of responses, hits, and false alarms. Participants responded significantly more often ( $M = 20.1$ ) than there were true changes (13) in each scenario booklet,  $W = 1958$ ,  $p < .001$ ,  $r = .79$ . The distribution of responses is plotted in Figure 37.

According to the “likely” change scoring rule (see Section 5.2.8), Participants on average hit 8.7 ( $SD = 2.7$ ) of 13 changes (67%) in each five-scenario booklet. However, at the same time they

committed 11.4 false alarms ( $SD = 5.2$ ). Diagnosis performance varied more between interface conditions and is discussed below in terms of diagnosis measures outlined in Figure 30, Table 16, and Table 17.

#### 5.5.3.1. Detection Interface Effects

The interface condition had no practically significant effect on hit and false alarm rates. Participants responded on average 20.2 times ( $SD = 6.6$ ) per 5-scenario booklet in the C condition, not significantly different from the 20.1 times ( $SD = 6.5$ ) in the C+R condition  $W = 296, p = .55, r = .02$ . In the C condition they hit on average 8.8 of thirteen changes ( $SD = 2.1$ ), no better than the 8.7 hits ( $SD = 2.0$ ) in the C+R condition  $W = 279, p = .78, r = .10$ .



**Figure 37 - Detection and Diagnosis summary, according to “Likely” scoring rule, aggregated by CUSUM-only or CUSUM+RE experimental condition. Top row are counts of R, AD, H, RD. Bottom row are rate-normalized according to Table 17. \*\* =  $p < .005$ , \*\*\* =  $p < .001$**

### 5.5.3.2. Diagnosis Interface Effects

As discussed above, participants responded frequently. They also attempted many diagnoses (AD, at top-center of Figure 37), both in the C+R condition ( $n = 33$ ,  $M = 17.0$   $SD = 6.3$ ) and C condition ( $n = 33$ ,  $M = 15.4$ ,  $SD = 7.3$ ). A Wilcoxon paired signed rank test did not find this difference significant,  $W = 218$ ,  $p = .50$ ,  $r = .001$ . Considering instead the likelihood of attempting diagnosis for a given response (ADr), a mixed effects generalized binomial regression (Appendix D.4.2) found that participants had significantly greater ADr odds when using the C+R interface,  $OR = 1.9$ , 95%  $CI [1.4, 2.5]$ ,  $z = 4.07$ ,  $p < .001$ , or with responses in their first experimental booklet,  $OR = 1.6$ , 95%  $CI [1.2, 2.1]$ ,  $z = 3.01$ ,  $p = .003$ . The practical magnitude of this change is that ADr increased from roughly 76% (Second booklet, C condition) to 90% (First booklet, C+R condition). The increased propensity of participants to attempt a diagnosis using the C+R interface confirms that we must evaluate diagnosis performance in terms of probability of right diagnosis (RDr), not absolute numbers of right diagnoses (RD).

For detected hits where the participant chose to mark a diagnosis cause (HD), the observed average diagnosis performance (RDr) in the C condition (Figure 37) was 28% correct, not significantly different from chance (1/4, 25%) according to a Wilcoxon signed-rank test  $W = 252$ ,  $p = .27$ ,  $r = .11$ .

With the C+R interface, RDr improved to 41% (Figure 37). This performance change was evaluated with a mixed effects logistic regression model (see Appendix D.4.3). The best model selected by maximum likelihood ratio explained RDr with a fixed effect of Interface  $z = 2.8$ ,  $p = .005$ , and TrueScenario as a random effect. The model estimated that the C+R interface raised the odds of a diagnosis being right by  $OR = 1.8$ , 95%  $CI [1.2, 2.6]$ . Main effects of FirstBooklet, Order, and School were non-significant,  $z > 1.1$ ,  $p > .27$  and removed stepwise. Random effects of Participant and Order did not improve the model when forward-fit.

### 5.5.3.3. Overall Interface effects

Overall M&T performance depends on both detection and diagnosis. By assessing overall effects, we can tell whether, for example, changes in false alarm rate negate effects of diagnosis accuracy. In the C condition, participants on average correctly hit and diagnosed 1.8 ( $SD = 1.3$ ) of 13 true changes, while with the C+R interface this increased to 3.1 ( $SD = 1.4$ ). As a

percentage of their (roughly equal) number of responses, correct response rate (CRr) in the C condition (9%,  $n = 33$ ) nearly doubled in the C+R condition (16%,  $n = 33$ ) significantly better according to a Wilcoxon paired signed rank test  $W = 73.5, p < .001, r = .41$ .

These changes in CRr (Figure 37) can be explained with a mixed effect binomial logistic regression model, in terms of a main effect of Interface,  $z = 3.48, p < .001$ , with a random effect of TrueScenario (described in Appendix D.4.4). Main effects of First Booklet, Order, and School were stepwise removed as they did not have significant effects,  $z > 1.07, p > .29$ . The C+R interface improved the odds of a response being completely correct by  $OR = 1.8, 95\% CI [1.3, 2.6]$ , consistent with the observed diagnosis improvements. These overall effects vary within the experimental scenarios. By-scenario effects are reported in Appendix E.1.

#### 5.5.4. Stimulus Effects

To evaluate what properties of changes made them more or less detectable and diagnosable, we examined the experimental data with respect to each change (rather than by-response). While changes are not truly independent of each other because they were combined in scenarios, change properties were semi-counterbalanced (Table 13) and were considered independent for the purposes of this analysis.

Each of the five scenarios contained one to four changes, with varying properties: cause type, large or small size, leading location, and accumulated evidence (see Section 5.2.2). The twelve<sup>11</sup> changes in five scenarios are not sufficient, however, to completely cross the change properties (Table 13), and there is only one example of each combination of change properties (two copies when mirrored in **G**<sub>1..5</sub> and **G**<sub>5..10</sub> scenarios). Thus, interactions of change properties cannot be generalized and were not analyzed. Further, by-change analyses inherently ignore responses that are not matched to a scenario change (FA). Percentages are described as how often a change was hit as a proportion of how many times it was presented (pH), rather than the average percentage of changes each participant detected (Hr).

---

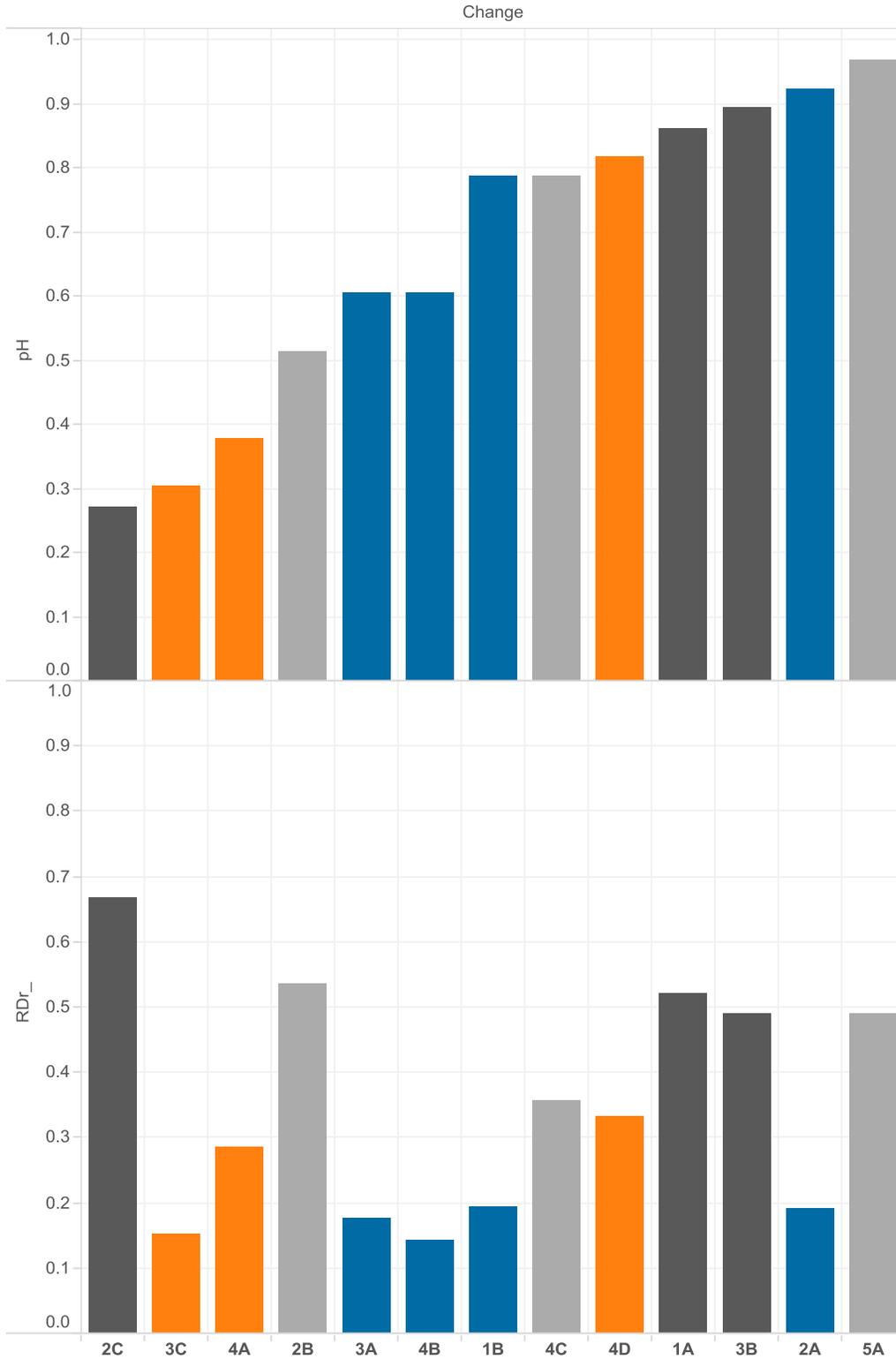
<sup>11</sup> thirteen if including Scenario 2's extra-large leading change 2A / 7A

#### 5.5.4.1. By-Change Properties Detection effects

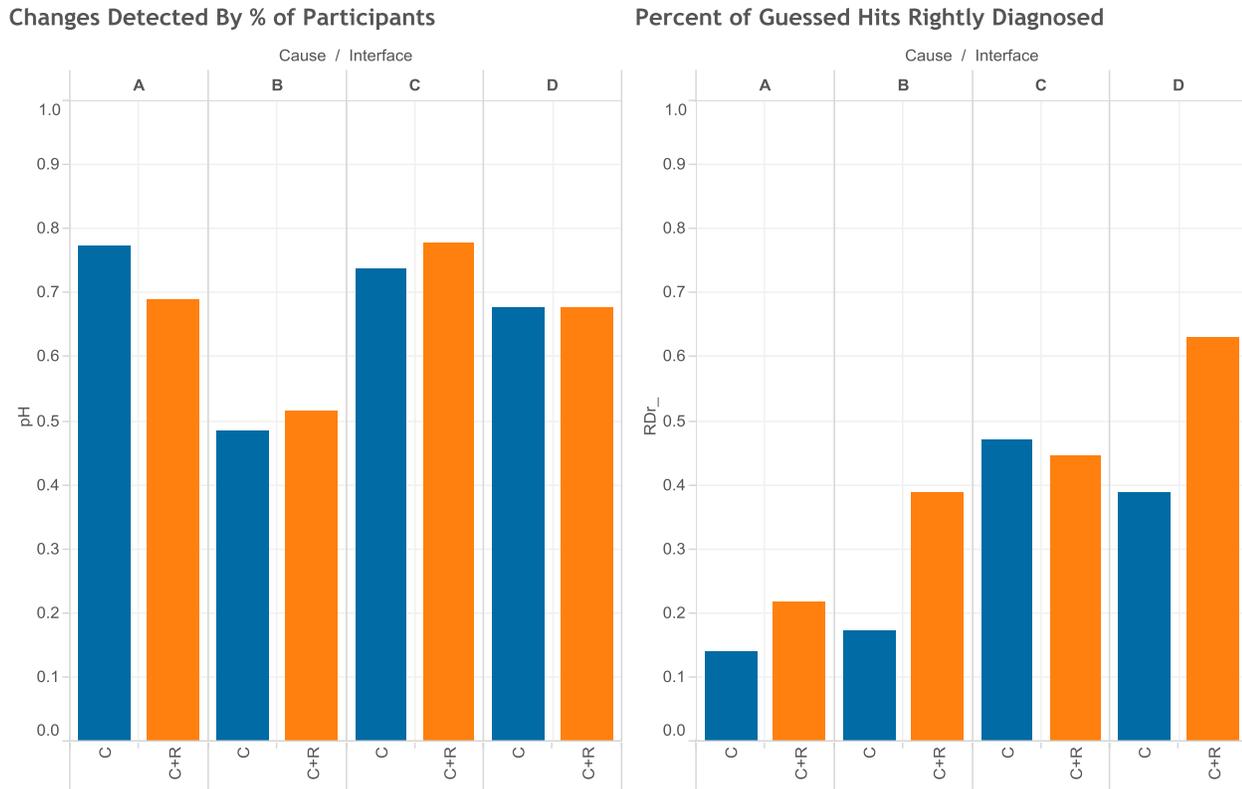
Each of 13 changes was detected by between 27% to 97% of participants (Figure 38). This variation was described by mixed effects logistic regression models, fit by stepwise removal (see Appendix D.6.1) to the 12 systematically-defined changes (omitting change 2A/7A that was used as a mask in Scenario 2/7). Maximum likelihood comparison indicated two candidate logit models.

The simpler model described detection with main effects of Large Sized changes  $OR: 6.2$ , 95%  $CI [3.9, 9.9]$   $z = 7.8$ ,  $p < .001$  and accumulated Evidence  $OR: 1.16 / \text{unit}$ , 95%  $CI [1.10, 1.21]$ ,  $z = 5.8$ ,  $p < .001$ , plus Participant and TrueScenario as random effects. A contrast of Evidence between the midpoint and the extremes presented in the scenario set (2.4 to 22.5 units of ‘average days consumption’), showed an  $OR$  of 4.29 / half-range, 95%  $CI [2.6, 7.0]$ . The more complex model retained SizeLarge,  $OR: 5.0$ , 95%  $CI [3.0, 8.5]$ ,  $z = 6.1$ ,  $p < .001$  and Evidence  $OR: 1.17 / \text{unit}$ , 95%  $CI [1.11, 1.24]$ ,  $z = 5.8$ ,  $p < .001$ , and also included Cause type  $|z| < 2.5$ ,  $p > .013$ , with only one Cause type significant. Main effects SizeLarge and Evidence were weakly correlated in both models,  $r < .27$ .

The effects of change Cause are summarized in Figure 39 and the details of model representation shown in Appendix D.6.1. Changes of cause type A (Baseload) were detected by 77% of participants, type C (Heating) by 74% of participants, and type D (Generator) by 68% of participants. Changes of type B (Workday) were detected by only 48% of participants  $OR: 0.52$ , 95%  $CI [0.32, 0.87]$ .



**Figure 38 - Detection and diagnosis rates of each of 13 changes (Labeled by scenario, 1A to 5A). Scored using "likely" rule, combining changes from Normal G<sub>1..5</sub> and Inverted G<sub>5..10</sub> scenario sets. At top, proportion of participants who detected each change (pH), at bottom right diagnosis rate (RDr). Color indicates change cause (Table 13, Blue = Baseload, Orange = HDD, Grey = Weekday, Black = Generator).**



**Figure 39 - Proportions of change Detections (pH) and associated Right Diagnosis rates (RDr), aggregated by change Cause/Type and Interface condition. Change Causes are A (Baseload), B (Workday), C (Heating), and D (Generator).**

### 5.5.4.2. By-Change Properties Diagnosis effects

Changes varied even more widely in their right diagnosis rates (RDr) (Figure 38), from 14% to 67% of detections and attempted diagnoses (HD) for each change. Consistent with the above diagnosis results by-participant and by-scenario, Interface condition seems to be associated with more right diagnoses (Figure 39). However, the analysis below should be interpreted with caution since it does not account for tendency to attempt each diagnosis cause (Section 5.5.5.6).

As in previous analyses, mixed effect logistic regression models were fit to describe RDr. Several candidate models were developed by reverse fitting, as described in Appendix D.6.2. The simplest mixed effects logistic regression model had main effects of Interface,  $OR = 1.7$ , 95%  $CI [1.2, 2.6]$ ,  $z = 2.62$ ,  $p = .009$ , and Cause,  $|z| > 1.88$ ,  $p < .06$  and should be more generalizable than more complex models with non-significant terms  $p > .05$ .

This simpler by-change RDr model quantifies the change cause/type effects on diagnosis rates in Figure 39 with reference to change type “A” (Intercept/Baseload). Changes to weekday

consumption (“B”) had greater odds of being rightly diagnosed, *OR*: 1.9, 95% *CI* [0.97, 3.5], as were Heating factor changes (“C”), *OR* 3.8, 95% *CI* [2.2, 6.5] and changes to the simulated Generator factor (“D”), *OR* 5.0, 95% *CI* [2.8, 8.8]. Other factors may influence these results, such as propensity to attempt diagnosis for each change cause (Section 5.5.5.6). This and other influences will be examined in the next section, before discussing the results in Section 5.6.

### 5.5.5. Other influences on Experiment II results

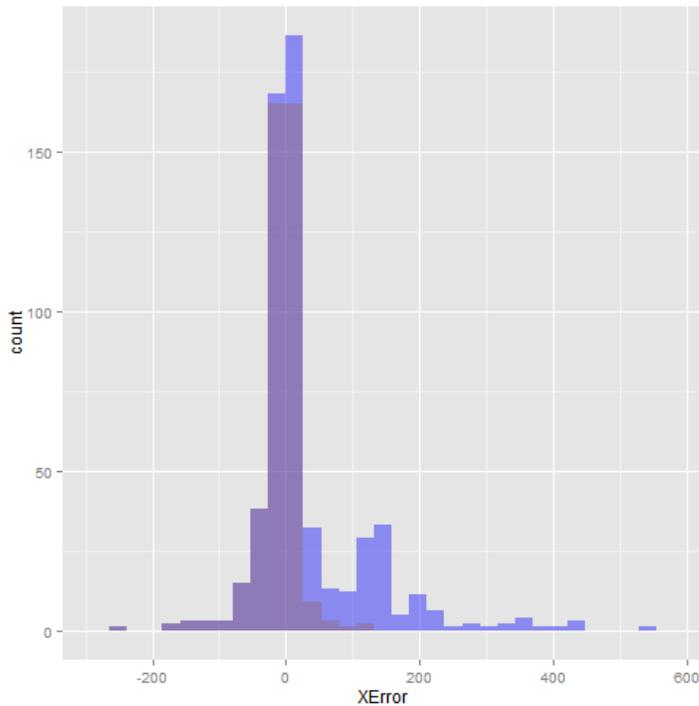
Besides interface, scenarios, and change types, other influences that could have affected the observed CUSUM and RE chart interpretation include data entry errors, the choice of data scoring rules, learning effects, use of perceptual aids, participant confidence, or participants’ tendency to attempt diagnosis.

#### 5.5.5.1. Data Entry validation

Errors in transcribing written and drawn experimental responses could affect results. To check how reliably the interpretation rules in Section 5.2.7 were applied, a 10% random sample of participant responses (127 samples) was re-entered from paper forms by two independent graduate student validators. Inter-rater reliability was calculated using Krippendorff’s Alpha (Krippendorff, 2004) for response Date, Confidence, Direction, Cause, as well as chart measurements such as location of indicated change, and exact distance of line marks. Scores exceeded  $\alpha > .95$  for all measures. The 20 largest mismatches were re-inspected and 17 were due to validator error, 3 responses were ambiguous, and none were due to mis-transcribed data (see Appendix D.7.1). These findings suggest that participant responses were accurately recorded.

#### 5.5.5.2. Scoring rules

The two scoring rules used to match responses to answers (Section 5.2.8) classified different numbers of Hits. Under the more conservative “InBox” scoring rule, 410 responses were scored as hits, while the more liberal “Likely” rule scored 576 (Appendix D.7.1). As shown in Figure 40, the “Likely” scoring rule forgave some late responses.



**Figure 40 - Histogram comparing response- change pairs scored by "InBox" and "Likely" rules, according to discrepancy between marked "X" and the true date of the matched change (in scenario days, 5 days = ~1mm). Darker shaded sections at left were scored by both rules. Lighter shaded scores at right are late responses only matched by "Likely" rule.**

While the experimental intent justifies use of the “Likely” rule (see Section 5.2.8), it is worth checking how consistently the scoring rules capture participants’ individual differences. The within-participant total Hits were strongly correlated across both scoring rules,  $r(64) = .72, p < .001$ , with a Chronbach’s Alpha of  $\alpha = .84$ . Similarly, by-participant Right Diagnoses were strongly correlated across scoring rules,  $r(64) = .82, p < .001$ , with  $\alpha = .90$ . Interface condition had no practically significant effect on mean differences between scoring rules. In both conditions, the InBox scoring rule marked 31% of responses as Hits and the Likely rule marked 43% (see Appendix D.2.2). There is little evidence that choice of scoring rule affects conclusions about interface effectiveness.

### 5.5.5.3. Learning effects

Participants took less time to complete their second experimental booklet (Section 5.5.2.3). However, this learning effect was not present in other measures of task performance. The main effect of booklet order did not significantly explain probability of hits (Appendix D.4.1 and D.6.1), probability of right diagnoses (Appendix D.4.3 and D.6.2), or overall probability of

correct responses (Appendix D.4.4 and D.5.4). First Booklet was only significantly associated with the likelihood of attempting diagnosis (Appendix D.4.2).

#### 5.5.5.4. Marked visual aids

Participants were provided with a ruler and pencil to annotate charts as they liked, as well as cardboard masks to visually isolate chart segments (as proposed in Section 4.3.9). The training briefing demonstrated how to draw vertical lines to link CUSUM and RE charts when they suspected changes. Since the experiment did not control or manipulate these aids, we cannot determine causal effects. However, to explore Hypothesis 4 (Section 5.1.3), each participant response was coded as whether it was accompanied by a vertical linking line. Of 1329 responses, participants marked a linking line on 68%. This varied slightly with interface condition, 72% of 667 responses in the C condition and 65% of 662 responses in the C+R condition.

In the C+R condition, where a benefit was hypothesized, an accompanying linking line had little practically significant correlation with whether a response was a hit or false alarm (FAR),  $\chi^2(1, N = 662) = 1.64, p = .20, \phi = .05$  or whether a hit with diagnosis was correctly diagnosed (RDR),  $\chi^2(1, N = 251) = 0.040, p = .85, \phi = .01$ .

#### 5.5.5.5. Confidence in responses

The confidence that participants assigned to each response had little association with whether the response Hit  $r(1316) = .13, p < .001$  or Rightly diagnosed  $r(462) = .02, p = .63$  a true change (Appendix D.7.3).

#### 5.5.5.6. Propensity to attempt diagnosis with each change type

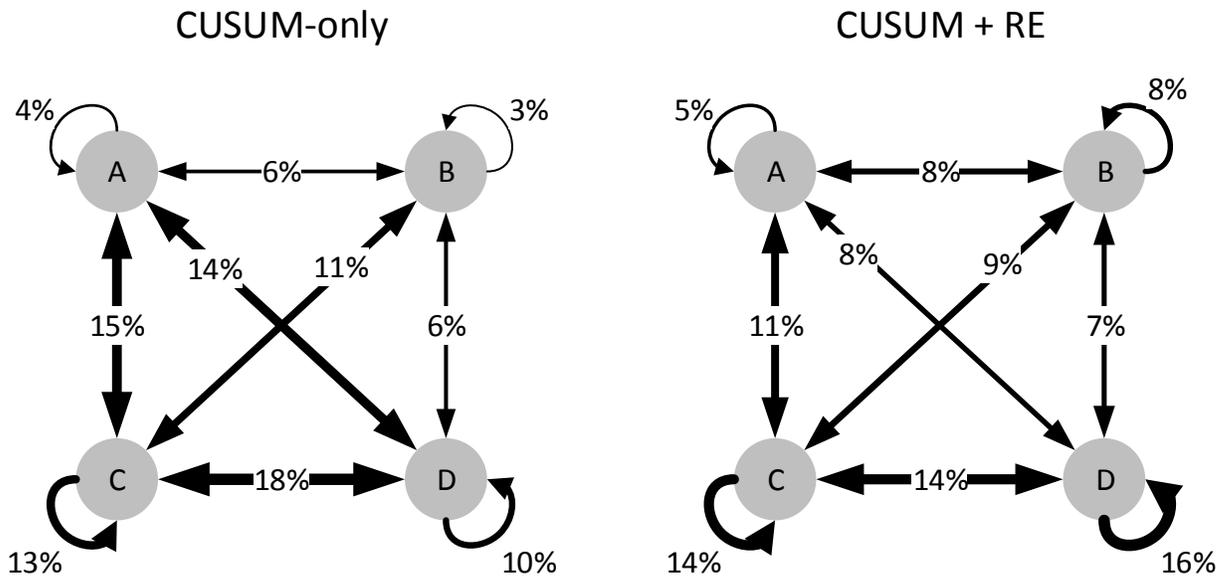
While the above influences on responses were not found practically significant, the propensity to attempt a diagnosis each change type may affect the diagnosis effects reported in Section 5.5.4.2. As shown in Figure 39, the likelihood of a participant detecting a change varied with change type more than Interface condition (see Appendix D.7.4.1). The likelihood of a detected change being correctly diagnosed varied even more between change types. However, the diagnosis analysis of Section 5.5.4.2 did not account for effects of participants attempting each diagnosis cause. For example, if a participant only ever diagnosed cause “C”, changes of type C would be correctly diagnosed whenever they were detected (at the expense of other change types).

This effect is worth investigating. Proportions of diagnosis attempts made by all participants in each interface condition differed from the (roughly) equal proportion of four change types in the scenarios  $\chi^2(3, N > 487) > 34.6, p < .001, \phi > .25$ , and differed between interface conditions  $\chi^2(3, N = 1024) = 28.9, p < .001, \phi = .17$ . Participants responded that changes were of type “C” (Heating) or “D” (Generator) more frequently than the true proportion present in the five scenarios. Cross-tabulations of attempted diagnoses against true cause type are shown below in Table 22. As an aside, while the scenarios contained four changes of type “A” (Baseload), rather than three as in each of the other types, Change 2A/7A was omitted in this analysis because it was the only extra-large-sized change, and served as an obscuring mask for Scenario 2 / 7 (see Section 5.2.2).

**Table 22 - Response counts, cross-tabulated by Marked Cause (including non-diagnosis-attempts N/A), by Interface condition, and by True Cause of Hit according to “Likely” scoring rule (plus False Alarms N/A). Extra-large 13th change 2A / 7A omitted. Bolded diagonals represent Correct Responses. Totals not shown.**

Cause marked with Attempt	TrueCause of Hit										
	With C-only Interface					N/A (FA)	With C+R Interface				N/A (FA)
	A	B	C	D	A		B	C	D		
A	<b>8</b>	4	9	6	41	<b>11</b>	4	11	4	65	
B	7	<b>6</b>	7	2	46	15	<b>17</b>	9	5	66	
C	20	15	<b>25</b>	22	126	14	12	<b>32</b>	12	76	
D	22	10	12	<b>19</b>	80	13	11	20	<b>36</b>	104	
N/A (no diag.)	14	13	20	18	84	8	7	5	10	65	

To examine attempted diagnosis behavior, we use Hits with Diagnosis (HD), the intersection of the 41% of responses that were hits and the 81% of responses where a participant attempted diagnosis. Diagnosis confusions (the non-diagonals of HD in Table 22) will be first summarized in terms of total responses, then attempted diagnosis cause. In terms of total responses, combining cross confusions (e.g. summing the 4 times A was confused with B and the 7 times B with A in Table 22), the confusion relationships can be diagrammed as in Figure 41. This shows the proportion of HD that were either correctly diagnosed, or confused with another change type.



**Figure 41 – For all hits with an attempted diagnosis (HD), confusions between attempted diagnosis and true change causes, as a percentage of total HD. Outside loops indicate right diagnoses (RD). Perfect performance would be 25% of HD being right diagnoses of each change type. Based on data from Table 22.**

Examining Figure 41, the most common confusions in both interface conditions are of change type C (Heating), confused with D (18%, 14%), and C confused with A (15%, 11%). In the C+R interface condition the most reduced confusion is between A and D (14% reduced to 8%). The proportionately most improved right diagnosis in the C+R interface condition is B (3% improved to 8%), though B is still confused about as often as before. However, these confusion metrics still do not account for how often participants attempt diagnosis with each change cause.

Confusions in terms of the proportion of HD with each change type attempted (pHD) are tabulated in Appendix D.7.4.2 and Table 23. This distinguishes the direction of confusions (e.g. Response of A hitting change of type B counted separately from a B attempt hitting A). Table 23 omits false alarms, since the proportion of marked causes for HD (totals of marked causes “A” to “D” in Table 22) was not significantly different from false alarms with diagnosis FD (Table 16)  $\chi^2(3, N = 1024) = 4.78, p = .19, \phi = .07$  (See Appendix D.7.4.3).

**Table 23 - Proportion of Hits matched against True Causes of each change type. Same data as from Table 22. Change 2A/7A omitted. Total by-diagnosis proportions (HDs) of Marked Causes for Hits with Diagnosis summarized at right.**

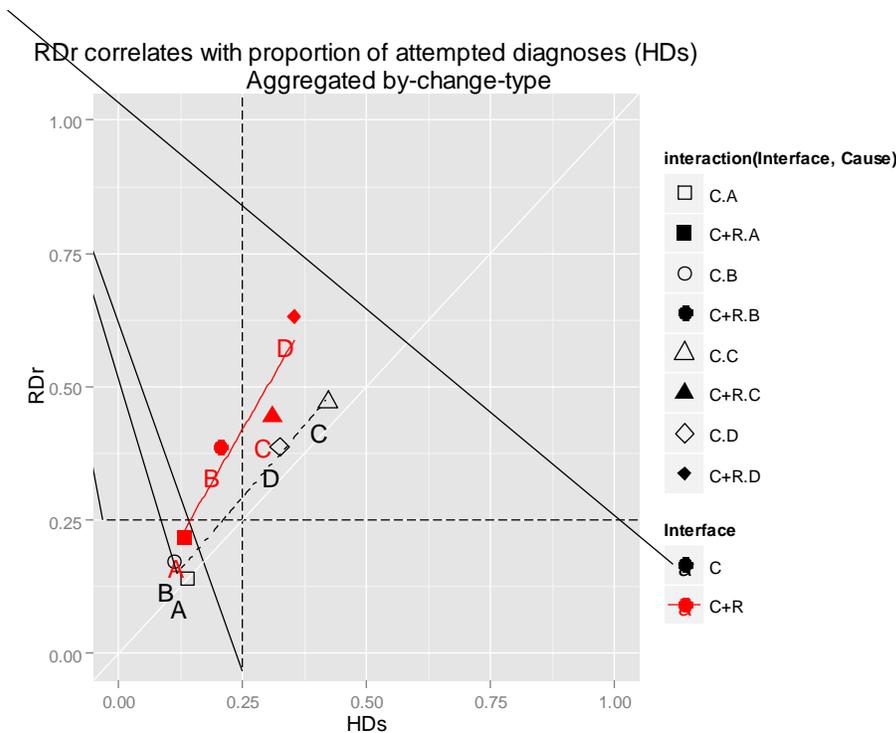
Marked Cause of Response	Proportion of Hits with Diagnoses matching TrueCause (pHD)								Split of Diagnosis attempts (HDs)	
	With C-only Interface				With C+R Interface				C only	C+R
	A	B	C	D	A	B	C	D		
A	<b>30%</b>	15%	33%	22%	<b>37%</b>	13%	37%	13%	14%	13%
B	32%	<b>27%</b>	32%	9%	33%	<b>37%</b>	20%	11%	11%	20%
C	24%	18%	<b>30%</b>	27%	20%	17%	<b>46%</b>	17%	42%	31%
D	35%	16%	19%	<b>30%</b>	16%	14%	25%	<b>45%</b>	32%	35%

As shown in Table 23, confusion between change types were not always symmetrical, as conflated in Figure 41. For example, a higher proportion of responses diagnosed as B hit changes of type A in each experimental condition (32%, 33%) than A responses indicated a change of type B (15%, 13%).

Within cause response types, the proportion of HD that were right (pRD) ranged from 27 to 30% of diagnosis attempts (HDs) in the C interface condition, and between 37 and 46% in the C+R interface condition. pRD for each cause type differ from the right diagnosis rates (RDr) analyzed in Section 5.5.4.2, because participants on average attempted diagnosis with each cause in unequal proportions (HDs in Table 23, right). The percentage of rightly diagnosed hits (RDr in Figure 39) is strongly correlated with the proportions of attempted diagnoses in hits (HDs) (in Table 23),  $r(4) > .98$  (see Table 24 and Figure 42).

**Table 24 - Proportion of Rightly Diagnosed hits with diagnosis (RDr) compared with tendency to attempt each change diagnosis type (HDs) accompanying a hit. Summarizes Figure 39 and plotted in Figure 42.**

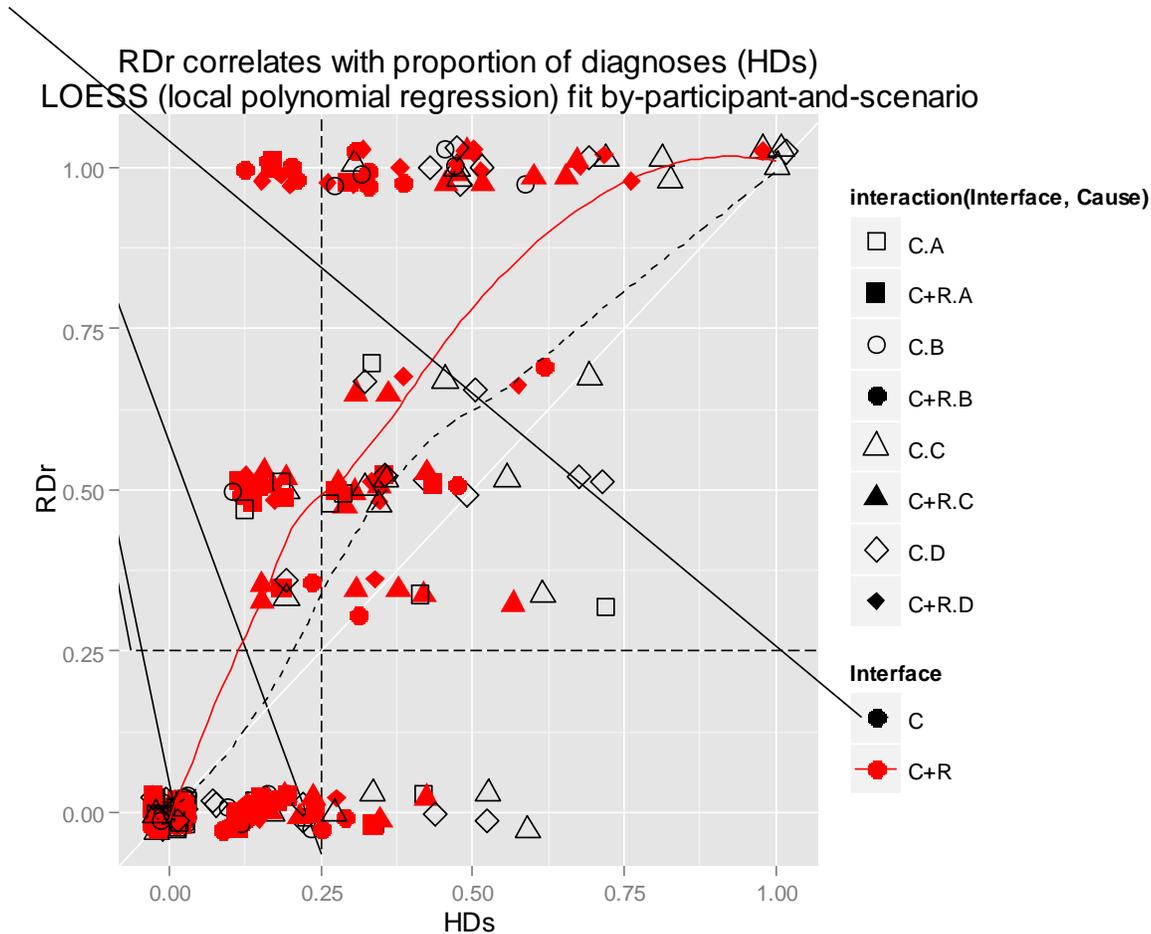
Cause	Interface			
	C		C+R	
	RDr	HDs	RDr	HDs
A	14%	14%	22%	13%
B	17%	11%	39%	20%
C	47%	42%	44%	31%
D	39%	32%	63%	35%



**Figure 42 - For Hits, how often change types were rightly diagnosed (RDr) correlates with the by-cause proportion of diagnosis attempts (HDs). Data aggregated by type of change Hit by response (A,B,C,D). Scale lines indicate chance (1 in 4) performance.**

The correlation, aggregated by change-type in Figure 42, shows that deviations in by-change diagnosis performance (Figure 39) from the observed average RDr (Figure 37) are associated with changes in tendency to attempt each diagnosis alternative. For example, while participants

achieved similar ~45% RDr for change type “C” (Heating) in both interface conditions, this required 11% more diagnosis attempts of type “C” in the CUSUM-only interface condition (with corresponding fewer diagnosis attempts for change type B).



**Figure 43 -** For each change type Hit, the percentage participants ( $n=33$ ) rightly diagnose (RDr) in a scenario booklet ( $n=264$ ) correlates with the proportion of diagnosis attempts (HDs). Scale lines indicate chance (1 in 4) performance. Data points are jittered to show density. Local least-squares trend lines fit by LOESS.

The correlation between diagnosis attempts and success is consistent even in less-aggregated data, for both C-only interface  $r(108) = .76$  and C+R condition  $r(121) = .65$  (see Appendix D.7.4.4). This is shown in Figure 43 as a trade-off between diagnosis commitment (HDs) and diagnosis success (RDr), analogous to a Receiver Operating Characteristic (ROC) curve in SDT<sup>12</sup> (Green & Swets, 1966). One way to quantify detection sensitivity is the area under the

<sup>12</sup> ROC curves plot probability of FA versus probability of H in single-stimulus trials. Chance performance is similarly defined as a straight line between 0% and 100% of both FA and H. However, this chart differs from ROC curves as it is not defined for HDs=0, RDr >0 (without attempting diagnoses, cannot be correct), and samples are not independent ( $\sum HD_{S_{A..D}} = 1$  for each participant-booklet).

ROC curve, where  $P(A) = .5$  represents chance performance. Applying this principle and integrating under the piecewise LOESS curve in Figure 43 calculates a ‘P(A)-like’ diagnosis sensitivity measure of .56 in the “C” condition, and .70 in the “C+R” condition.

Another perspective on Figure 43 is that participants in the C+R interface condition on average committed 10 to 20% fewer HDs per change category to achieve comparable RDr. Figure 43 also shows that participants strongly committed to one change category (HDs > 50%) more often in the C-only interface condition ( $n = 18$ ) vs. the C+R condition ( $n = 11$ ). This analysis cannot conclusively determine if the C+R interface particularly helped diagnose certain change types. It suggests that tendency to attempt diagnoses accounts for much of the variability, and that the C+R condition was associated with more equal diagnosis attempts.

### 5.5.6. Participant Feedback

Because the experiment was conducted in a group setting, participant feedback was solicited in written form.

#### 5.5.6.1. Questionnaire

Participants responded that the RE charts better showed what type of changes happened ( $n = 32$ )  $W = 39.5$ ,  $p = .002$ ,  $r = .35$ , and were less confusing ( $n = 33$ ),  $W = 118$ ,  $p = .008$ ,  $r = .30$  according to Wilcoxon matched-pairs signed rank tests. Questions on ease of use, timing, and informativeness showed little difference between interface conditions,  $p > .53$ . Summary statistics are shown in Table 25.

**Table 25 Questionnaire Summary Statistics, for all participants (n=33). Score of 3 is "Un-decided".**

<i>Likert Scale</i>	<b>Question asked about</b>			
	CUSUM chart		RE chart	
	Mean	SD	Mean	SD
<i>1 = Strongly Disagree</i>				
<i>5 = Strongly Agree</i>				
<i>Were easy to understand</i>	3.6	1.0	3.7	0.9
<i>Showed when changes happened</i>	3.7	1.0	3.8	0.8
<i>Showed what type of changes happened</i>	2.8***	1.0	3.7***	0.9
<i>Were informative</i>	3.8	0.8	3.9	0.6
<i>Were confusing</i>	2.7**	0.9	2.3**	0.8

Correlations between questionnaire responses are described in Appendix D.7.5.1.

### 5.5.6.2. Comments

Participants were also asked if they had any other comments about the experiment. Comments on the experimental method or apparatus consisted of:

*“Parameter charts were much easier to interpret”*

*Seems good idea, “maybe add a course of action” in future revisions?*

*Seems useful when realtime (building automation) system data logging is not available*

*Need “back-up information on what this whole thing is really about”*

*“The black ‘windows’ helped out a lot”*

*“Ability to draw vertical lines ... will impact readings more the more parameters are added.”*

*Difficult to distinguish actual/predicted utility consumption series.*

*“Too many variables, hard to distinguish, except for obvious changes”*

*“A little more background information for participants”*

*“The CUSUM charts clearly outlined excessive or reduce use in energy with the aid of parameter charts and energy use”*

## 5.6 Experiment Discussion

To discuss the results of this experiment, we’ll address the motivating principles (Section 5.1) and the hypotheses (Section 5.1.3) in turn.

### 5.6.1. Evaluating performance with standard (CUSUM) tools

Using the CUSUM-only tools, participants detected on average 68% of thirteen changes per scenario booklet, with wide variation - the worst detected only 1/3 and the best detected all. This hit rate is encouraging, except that the average false alarm rate was 57%, meaning that more than half of responses did not indicate a true change or duplicated an already-detected change. False alarms were common and most participants committed between 50% and 62% FAR.

The high false alarm rate is concerning. The worst-performing participant indicated 2.8 times as many responses as true changes present! Such behavior in a practical work situation would not foster credibility or trust (see Section 2.5.9). Interestingly, this behavior directly contradicted field study observations (Section 2.5.1) where participants seemed reluctant to commit to identifying changes. Participants may have believed that transient chart variations indicated practical concerns, since the experimental scenarios did not give participants any referent to judge the practical (financial) significance of CUSUM chart displacement. Chart roughness may have affected false alarms: Scenario 2, whose CUSUM chart was smoother due to the large initial change, had the lowest false alarm rate, 30% (discussed below in Section 5.6.3 and Appendix E.1.1). False alarms indicate over-confidence, as participants were instructed to only mark responses they felt confident of, and the indicated confidence levels showed only a small ( $r = .1$ ) correlation with detection performance. This overconfidence may be explained by participants being college students with limited experience to calibrate their sense of their own abilities, and not having received individual feedback about their detection performance.

Diagnosis with CUSUM-only charts was no better than chance (RDr one in four) performance. Given that CUSUM charts provided at least the diagnostic cue of a “scalloped” chart shape for

change type C / Heating (Figure 5 in Section 2.3.5), this suggests that participants did not perceive CUSUM diagnostic cues, or that participants attended to misleading cues. Whichever the case, participants attempted diagnoses frequently. Despite instructions to only mark a cause when sure, on average participants attempted diagnosis (ADr) for three out of four responses. Detection and diagnosis accuracy were not associated. The combined effect of a high false alarm rate and chance-level diagnosis performance was that on average only 9% of responses participants made in the CUSUM-only condition correctly identified a change (CDr) and only 14% of changes present were correctly detected and diagnosed (Hr).

These absolute performance measures should be interpreted with caution, as only one set of scenarios was used, participants received feedback in only three practice problems, and participants had no access to naturalistic cues, or tools to support other identification and diagnosis strategies (Section 3.5). However, the results do indicate that time series charts of consumption, control, CUSUM, and driver variables are not a sufficient representation to support effective M&T diagnosis.

### 5.6.2. Evaluating support of Recursive Estimates-based strategies

In trials where participants could consult RE charts to supplement CUSUM charts, overall performance significantly improved. However, this was not due to improved detection. As hypothesized, the CUSUM+RE condition caused no practically (or statistically) significant difference in participants' average response profligacy (20.1 responses per booklet vs. 20.2), hit rate (68% vs. 67%), or false alarm rate (57% vs. 57%). This is consistent with participants appropriately relying on the CUSUM charts for detection (Figure 35), rather than false alarm-inducing RE chart variations (due to endogeneity and autocorrelation, see Section 4.3.4). This is also consistent with briefing instructions.

However RE charts were hypothesized to *reduce* false alarms based on to the short-timescale (light grey) chart shifting for new changes but not for recurring changes (see Section 4.3.8). This effect was not observed, possibly due to participants' tendency to over respond, or due to the short-timescale charts not forming perceptible characteristic "square wave" shapes demonstrated in the experimental briefing.

Diagnosis behavior changed significantly with the RE interface. Participants attempted diagnosis (ADr) significantly more often (85% vs. 76%), independent of whether the response hit a change or not. Magnifying this increased ADr, participants' diagnoses of hits were right (RDr) on average 41% of the time, significantly more often than expected by chance (25%) and observed in the CUSUM-only condition (28%). Overall, the improved diagnosis performance significantly raised the average proportion of completely correct responses (CRr) to 15% in the RE+CUSUM condition (from 9%), and the proportion of the thirteen changes present correctly detected and diagnosed to 24% (from 14%).

Participants were familiar with CUSUM chart interpretation from their college program courses. Despite the RE display being novel and participants only having learned the method from one briefing and three practice trials, diagnosis performance improved without taking significantly longer to use RE charts. The average task pace of 40 minutes per booklet in the CUSUM-only condition increased only to 45 minutes in the CUSUM+RE condition, which was not practically or statistically significant. The 5 minute increase is consistent with participants simply having consulted the RE charts. In practical M&T work, correctness would be more important than speed. Differences in well-practiced performance would be more practically relevant but were not evaluated in this experiment. Overall, the RE chart supplement caused more effective M&T diagnosis behavior.

### 5.6.3. Scenario effects on detection and diagnosis

Scenario effects were considered in Appendix E.1. Response rates and corresponding false alarm rates varied between scenarios (Figure 64 and Figure 65). The fewest responses were observed in Scenario 2, where participants responded about half as often in other scenarios (Section 5.2.2 and Appendix F.3). Scenario 2 was inspired by use of Measurement and Verification (ASHRAE Guideline Project Committee 14P, 2002) models for M&T, and was developed with a large, early change to create a steep CUSUM slope, expand the chart scale, and thereby obscure later changes (as in Figure 8). This had the side effect of making the CUSUM chart appear smoother. It is not clear whether CUSUM chart smoothness or the “straight line” shape account for participants marking fewer responses. Later changes seem to have been obscured, though: changes 2B/7B and 2C/7C were the least and 4<sup>th</sup>-least detected of 13 changes (Figure 38). RE charts had no significant effect on tendency to respond in any scenario.

The most false alarms were observed in Scenario 5, which attracted roughly the same number of responses as Scenarios 3 and 4 despite containing only a single change. Scenario 5 was developed to test if participants would realize that the six CUSUM chart slope changes (e.g. Figure 5) were due to a single change in the heating parameter affecting winter-time utility consumption over three years. While participants detected Scenario 5's lone change 97% of the time (Figure 65 and Figure 38), its after-effects attracted just as many responses as the three and four changes of Scenarios 3 and 4, respectively. Contrary to expectations, RE charts had no significant effect in reducing false alarms in any scenario.

Scenarios with more changes generally had lower hit rates (Table 31). Hit rates were highest (97%) for the scenario with one change (Scenario 5), moderate for the scenario with two changes (Scenario 1), and comparable (57-65%) for the remaining scenarios with three or four changes. Scenarios with more changes did not necessarily have worse diagnosis rates, however. Variation between scenarios and interfaces (Figure 65) was significant, but interaction effects found only weak evidence ( $p > .03$ , uncorrected) for scenario-specific interface effects.

#### 5.6.4. Change type effects on detection and diagnosis

Each scenario contained between 1 and 4 changes, with properties crafted to represent permutations of leading/trailing, large/small, and cause type, and accumulated evidence (see Table 13). This is not comprehensive; since the experiment only had one scenario set, change properties were confounded with emergent effects of the particular scenario arrangement. Because of this, evidence from this experiment is not enough to definitely infer property effects on change detection and diagnosis. The intent of exploratory data analysis was to suggest whether properties of each change such as leading/trailing (ambiguity), size (chart slope), type (chart shape), or accumulated evidence (vertical displacement) are more or less usefully represented on CUSUM or RE charts.

The RE Interface did not mediate any change properties' effects on detection. Data supported two models of change detectability (Section 5.5.4.1), the simplest of which explained detection in terms of Size and Accumulated Evidence<sup>13</sup>. The simpler model, all other things being equal, showed a change of twice the size (a 20% parameter change versus 10%) had significantly higher

---

<sup>13</sup> The more complicated detection model, not applied, included a main effect of change type, though only change type B (to weekday-only consumption) was significantly less likely to be detected.

odds of being detected by a ratio of 6.2 to one. At the same time, the model estimated a change with a mid-scale accumulated evidence would have 4.3 times greater odds of being detected than a change with the lowest level of evidence (e.g. 12.4 ‘average day consumption’ units vs. 2.4) present in the experimental scenarios<sup>14</sup>. The scenario set was crafted so that size and evidence were weakly correlated ( $r = .18$ ). This allows interpreting the results to mean that CUSUM slope changes (size) and vertical displacement (evidence) are complimentary cues for change detection. The lack of a significant Interface interaction shows that as expected, RE chart cues were either not useful, ignored, or misinterpreted for detection of all change types.

Changes of different cause types were diagnosed at practically and significantly different success rates from 14% to 67% (Section 5.5.4.2 and RDr at bottom, Figure 38), depending on the Interface used. No other properties of changes (size, leading or following, or accumulated evidence) had a statistically significant effect ( $p < .05$ ) on RDr. The non-significant Size effect suggests whichever chart cues people used to diagnose were sufficiently perceivable in changes noticeable enough to be detected. It is not clear, however, what cues or mechanisms account for different RDr for each change type.

An influence which correlates with RDr is how often participants attempt diagnosis with each change type (HDs) within interface conditions (Section 5.5.5.4 and Figure 42). Other influences of change type can be partially isolated by using normalized proportion-of-HD measures in Table 23, which show that in the C+R condition change types C and D had the highest proportion of HD being RD (46% and 45% pHD). However, since AD, Hr and RD are confounded by response bias, it is difficult to draw statistical inference. While changes to heating and gas generator (types C and D) were more correctly diagnosed, it is not clear how much varying RDr between change types is due to the resulting chart cues being more distinctive (in the Table 23 C+R condition), or if it is mostly explained by participants indicating that change type more frequently (Figure 39). This experiment could not distinguish influence on diagnosis attempt rates from:

- Chart cues, the distinctive shapes of change types C (heating) and D (generator), as hypothesized
- Participants’ prior expectation of how often each change cause was present (none was given in the instructions or briefing)

---

<sup>14</sup> Uninterrupted time between changes was evaluated post-hoc for predictive value, but was not well-counterbalanced in the scenario set, was undesirably correlated with Large Size, and had a weaker, non-significant relationship with detection.

- The phrasing of each causal explanation (C “Heating” or D “Generator” may have seemed more diagnostic, plausible, or familiar)
- The order in which RE charts and diagnosis forms were printed (from top to bottom D,C,B,A, see Appendix F.5).

Change type “A” (baseload) was guessed least often (Figure 38), perhaps because it appeared last on the list, or its phrasing as “always-on, everyday” seemed less specific and vivid (Tversky & Kahneman, 1973). This result is discussed further in Section 5.6.6 below.

To summarize, changes that were larger and produced more evidence (size \* uninterrupted duration) were more likely to be detected, regardless of the interface used. While some change types were more likely than others to be rightly diagnosed, this may reflect participants’ tendency to attempt (or guess) particular change diagnoses. Access to RE charts on average improved participants’ right diagnosis rate for all change types, regardless of their tendency to attempt diagnosis of that change type.

### 5.6.5. Support for experimental hypotheses

Only one of the five guiding hypotheses (Section 5.1.3) was confirmed by the results:

- 1) RE charts will have no effect on detecting changes.

This hypothesis is consistent with the (lack of) evidence. There was no practically or statistically significant difference in hit rate between interfaces (Sections 5.5.3.1 and 5.6.2).

- 2) RE charts will help participants commit fewer 'false alarms' due to misidentifying recurrent effects of persistent changes.

This hypothesis was not supported. There was no practically or statistically significant difference in false alarm rate between interface conditions (Sections 5.5.3.1 and 5.6.2), even in Scenario 5, which was specifically crafted to contain mostly recurrent effects of one persistent change (Section 5.5.3.2 and 5.6.3).

- 3) RE charts will improve change diagnosis of non-seasonal ‘driver’ variables, because CUSUM charts less clearly distinguish them from Intercept/Baseload changes.

This hypothesis was partially supported. RE charts significantly improved change diagnosis performance on average (Sections 5.5.3.1 and 5.6.2), except for changes to seasonal heating factor (type “C”). (Section 5.5.4.2 or Figure 39). However, diagnosis performance interaction effects of interface and non-seasonal ‘driver’ variables Workday (“B”) and Generator (“D”) were only marginally significant (Section 5.5.4.2). Differences in diagnosis performance within interface conditions may have been due to participants’ propensity to indicate diagnosis causes (Section 5.5.5.5).

- 4) Participants who draw linking lines between CUSUM charts and RE charts to diagnose will make fewer false alarms and more correct diagnoses.

This hypothesis was not supported. Since participants did not always draw linking lines, this hypothesis had to be evaluated at the per-response level. Whether a response was connected with a linking line had no practically or statistically significant association with whether the response was a false alarm. Similarly, linking lines were not practically or statistically associated with right diagnosis of hits (Section 5.5.5.4). This experiment could not measure whether participants used provided rulers or cardboard masks as visual aids alone without drawing linking lines (as suggested by one participant in Section 5.5.6.2).

- 5) Change magnitude will be a better predictor of detection and diagnosis than accumulated change evidence.

This hypothesis was partially supported. Change magnitude and accumulated evidence were the two properties significantly associated with whether a change was detected (Sections 5.5.4.1 and 5.6.4). However, their effects were comparable to each other within the range of stimuli in the experimental scenarios, and neither was significantly associated with diagnosis performance (Section 5.5.4.2).

### 5.6.6. Is the experimental validity sufficient?

This experiment was designed to be internally valid for drawing conclusions about effects of the CUSUM+RE interface. Block randomized counterbalancing of scenario and interface orders allowed control of learning effects. Learning effects were also mitigated by instructions and compensation rules encouraging participants to take as long as they needed. This seems to have

been achieved as learning effects were observed for time taken (Section 5.5.5.3) but not in detection or diagnosis performance measures. Data entry was validated for reliability (Section 5.5.5.1), and the chosen response scoring rule (Section 5.2.8) compared and justified (Section 5.5.5.2).

Experimental control of interface condition reduced the internal validity of conclusions about scenario effects (Sections 5.5.3.2 and E.1.1) and properties of individual changes (Section 5.5.4) because only one ordered set of underlying scenario changes was presented. Participants saw the same underlying scenario changes twice, once inverted and obscured by different random variation. However, no participants commented on this to the experimenter or in the questionnaire, and First Booklet effects were non-significant in all detection and diagnosis models.

The scenario set was also chosen as a compromise for ecological validity. Participants were recruited from two relevant, independent career training programs, so as to generalize to the energy management practitioner population. Participants from each career training program did not differ significantly in detection or diagnosis performance. Recruiting from career programs ensured that participants would be familiar with the ‘cover story’ of the hypothetical underlying system (a building with heating and on-site power generation). Recruiting sufficient numbers of such participants required the experiment to be administered on-site in a paper-based format, which required a fixed, uniformly-generated scenario set and limited the measures that could be taken. This lone scenario set, experimental control, response format, and scoring modes limit external validity:

- The underlying system weighted each energy driver to have equal contribution (Section 5.2.1) for internal validity, but is un-typical and may reduce the perceptual usefulness of the consumption charts by being too ‘jagged’.
- Model endogeneity (measurement error or un-measured variables) was present but well-behaved in the experimental scenarios – only mildly autocorrelated ( $\alpha=0.2$ ) normal measurement error and a Weibull distributed un-measured variable.
- Only one scenario set and the 13 combinations of change properties were tested. Fortunately the range of changes tested seems to have avoided edge effects (Figure 38) of zero or 100% detection or diagnosis.

- The random variation in the scenario stimuli was identical across every participant. This conflated random noise with the appearance of introduced changes. Re-generating random noise across every trial would eliminate this conflation.
- The changes introduced in this experiment were fairly large (10-20% shifts in energy performance) relative to the unexplained random variation (5-7%), compared to what can occur in practice.
- The information availability in this experiment (only the cover story and charts) was impoverished compared to what was observed in practice (where workers had at least knowledge of history and pre-existing suspicions of changes).
- Participants did not have to justify their conclusions to work colleagues or superiors, which would be expected to make them more cautious in marking changes.
- As suggested by the lower false alarm rate in the ‘smooth’ Scenario 2 (Figure 64), the auto-scaled formatting of the charts may have mediated participants’ sense of what change magnitude was practically significant<sup>15</sup>.

Not all experimental limitations harm external validity. Not being able to explain how participants chose to attempt diagnoses (Section 5.5.5.5) is less concerning to practical applications, since workers’ pre-existing beliefs and suspicion of likely diagnoses will be guided by other information sources (prior experience or conclusions from other diagnosis strategies). Similarly, subjects being novices and having only brief introduction to RE charts is not untypical of M&T in practice and is externally valid. And while practitioners may have access to software tools beyond pencil-and-paper, static charts are a lowest common denominator.

## 5.7 Conclusions for M&T detection and diagnosis

This experiment demonstrated a method to evaluate human performance at energy M&T using CUSUM chart and Recursive Estimates-based structural change detection strategies. While the experimental results are limited to the single set of synthetic energy data and five naturalistic scenarios tested here, detection and diagnosis performance measures developed in this experiment (Table 16 and Table 17) can be applied to future experimental or field work.

---

<sup>15</sup> Shown by short confidence windows (Section 5.5.2) and high false alarm rates (Section 5.5.3). Participants seem to respond mostly to surface features of chart.

Results showed that while changes were usually (68%) detected, this was at a cost of substantial false alarm responses (57%). Participants responded frequently, despite being instructed to only mark changes in which they were confident. This is inconsistent with the field study, which found participants reluctant to declare changes in energy performance. There are many factors that could explain participants' tendency to respond frequently in the pencil-and-paper task. Novice college students had no social pressure to be truly certain in detecting changes. They also had little information about typical variability of the synthetic energy-consuming system on which to base their expectations of background transient variability. A practice trial with zero changes might have helped calibrate participants' expectations. Participants seem to have simply responded to surface features of the CUSUM charts; Scenario 2 with the smoothest-appearing charts induced the fewest responses and both low hit rate and low false alarms. This suggests that visualizations for M&T should ensure that the visual scale of charts are calibrated to practical importance (e.g. a practically significant dollar amount). It also suggests that re-using Measurement and Verification (M&V) models for M&T will have a side effect of decreasing workers' tendency to respond.

Step changes to each of four "driver" variables were designed to produce different characteristic CUSUM chart shapes, but none were significantly more or less likely to be detected. Only change size and the accumulated evidence (change magnitude times uninterrupted duration) significantly affected how likely a change was to be detected. Doubling change size sextupled (6.2x) detection odds, and each extra 'average days consumption' of accumulated evidence comparably increased detection odds (1.2x). Changes that appeared first in the monitoring set were no more likely to be detected than changes later on, suggesting that "stale" M&T models remain useful for detection even after known changes in energy performance have occurred.

Diagnosis results showed that participants using only paper consumption, control, driver variable, and CUSUM charts did not diagnose at significantly greater than chance performance. This is consistent with field study observations of ineffective diagnosis. However, the experiment did not support every diagnostic tactic for using CUSUM charts, such as zooming in to better perceive day-to-day differences between modeled and actual consumption.

Contrary to expectations, changes that were leading, large, or had more accumulated evidence were no more likely to be successfully diagnosed in either interface condition. Only two levels of

change size were present in experimental scenarios, so this may reflect edge effects of diagnostic cues being easily perceptible even for ‘small’ (5%) changes. Change causes Heating and ‘Generator’ (driver parameters) were more often diagnosed, but this may simply reflect that participants attempted those diagnosis causes more often. For example, participants in the CUSUM-only condition chose change type “C” / Heating performance for 42% of their diagnosis attempts. In naturalistic behavior, participants’ expectation of diagnosis causes would be informed by richer contextual understanding of conditions and activities than this synthetic task supported.

The multiple-change structure of this experiment was chosen to be naturalistic, but as a side effect could not isolate causal effects of particular chart features on diagnosis or false alarm rates. There are many opportunities for future work, discussed in Chapter 6. However, the experiment provided evidence that Recursive Estimates charts caused significantly improved diagnosis performance, on average 41% correct (from 28% in the CUSUM-only condition, and 25% at chance performance). The RE charts were well-adopted (almost half of responses marked on them), did not significantly increase time taken (5 minutes, on average), and did not affect detection or false alarm performance. Participants indicated in questionnaires that they found RE charts significantly better at showing what type of changes happened, and less confusing. This was corroborated by participant comments. RE charts seem to induce effective M&T diagnosis strategies and should be considered in practical applications.



## Chapter 6

### Discussion and Future work

This dissertation investigated Monitoring & Targeting (M&T) energy end-use because of its societal importance in enabling effective energy management. Addressing climate change will require widespread improvement in utility energy efficiency through drastic increases in capital investment, combined with more sophisticated operation and maintenance of energy-intensive equipment. Hidden costs of energy efficiency are a major barrier to capital investment and energy control, even in contemporary heavy industry. Hidden costs comprise investigative and diagnostic labor of developing knowledge about site-specific utility energy efficiency. M&T software tools can be improved to support quicker, more effective problem solving by less expert labor, but which specific tool improvements are needed is unclear. Despite energy M&T having been practiced relatively unchanged since the 1980s, practical challenges are little understood since M&T literature is exclusively instructional and motivational rather than descriptive and reflective.

This dissertation proposes some observed barriers to effective M&T performance, characterizes them in engineering terms, develops, and evaluates a novel yet practical extension to M&T work support tools. The research process was development- and problem-driven, synthesizing field study and literature review through work analysis to controlled evaluation (Vicente, 2000). This entailed field study at two medium-sized energy consuming sites in Ontario, Canada, observing M&T practice by both remote energy analysts and on-site energy specialists. Work Analysis characterized field study findings in cognitive engineering terms and structured design requirements to enhance a popular cognitive strategy. Finally, a controlled experiment both described M&T performance metrics and quantified effects of the designed diagnosis aid.

Boundaries of the research narrowed at three points in the process. From an initial topic of Energy Management, the field study examined M&T in particular since it is a more general-purpose task, and a measurement channel to inform subsequent energy efficiency activities. While the work analysis identifies several strategies, only Comparative Analysis formed the design basis for a diagnosis support tool due to its observed popularity, its potential for versatile

use at many timescales or application areas, and tractability for controlled experimental evaluation. Finally, the experimental evaluation included only one synthetic task environment to maximize the experimental power of limited research participants in distinguishing performance with standard and extended diagnosis support tools. The focused research answers four main questions:

- What are some barriers to effective M&T experienced in naturalistic work settings? How do these difficulties contrast with M&T as described in practice literature?
- What is the design space for M&T cognitive work support tools? What constraints in the M&T work environment, task structure, and possible cognitive strategies are promising bases for design?
- How can cognitive engineering inform design of general-purpose M&T tools to address practice challenges?
- What features of M&T information environments affect task performance, and do work support tools change task performance as anticipated?

The dissertation describes difficulties in M&T practice under-discussed in the literature, particularly ineffective diagnosis and naive mental models. It presents a Cognitive Work Analysis which characterizes and contrasts M&T environmental structures, task information flows, and cognitive strategies. This analysis introduces novel framings of M&T as coupled representation-maintaining and problem-solving tasks, and diagnosis in terms of distinguishing ambiguity in the world and in representations. Guided by analysis conclusions, the dissertation describes how a promising contemporary statistical strategy extension was adapted using cognitive engineering principles to support diagnosis in the M&T work environment. Finally, it describes a novel controlled apparatus to simulate M&T problem-solving and quantify performance. An experimental evaluation of the designed M&T diagnosis aid substantiated some of its intended performance benefits and provided the first quantitative assessment of performance at an M&T task.

There are many opportunities for future work to better understand how to induce effective M&T behavior in global industry and beyond. The work described in this dissertation has contributed in four ways towards this goal.

## 6.1 Contributions

This dissertation has developed the following conclusions, arranged by its four main contributions to literature and practice.

### 6.1.1. Naturalistic description of M&T behavior

The descriptive field study of M&T work (Chapter 2) in this dissertation is a novel contribution to the M&T literature (Hilliard et al., 2014). As the first reported field study of naturalistic M&T behavior it compliments interview studies (Sandberg & Söderström, 2003), technology case studies (Capehart & Capehart, 2005), instructional literature (Carbon Trust, 2008; Gotel & Hale, 1989), and information system design guides (Efficiency New Brunswick, 2010). While this field study examines only a small convenience sample (five participants working in three environments on two business sites over eight weeks with one M&T information system) and suffers from sampling bias, it nevertheless contributes a data point to the existing (incomplete) description of M&T. In particular, new findings include:

- the pace of change in business equipment and processes causing models to fall out of date and create ambiguous CUSUM charts
- two additional characteristic cases where CUSUM charts are perceptually ambiguous due to stale models: large obscuring change, and multiple overlapping changes.
- a large fraction of worker time spent effortfully assessing data quality and determining the meaning of statistical models
- two different mis-understandings of energy performance models

I did not observe any energy-saving interventions during the eight week field study, which corroborates reported problems with M&T cost-effectiveness. Participants successfully confirmed effects of previous capital investments and known operation decisions (i.e. equipment failure, production planning). However, they had difficulty distinguishing effects of known changes from possible new effects and developing initial diagnoses of suspicious changes in energy performance. The M&T tool observed in the field study did not adequately support inspecting energy or driver data and presented very little information about statistical modeling. This is representative of M&T support specified in the literature (ASHRAE Guideline Project Committee 14P, 2002; Carbon Trust, 2008; Gotel & Hale, 1989) and I argue not adequate to

support effective diagnosis by workers who are not both fluent in statistical modeling and involved in day-to-day business operations.

### 6.1.2. Cognitive Work Analysis of M&T

Chapter 3 of this dissertation describes M&T in Cognitive Engineering terms through three CWA phases: Work Domain Analysis (WDA), Control Task Analysis (ConTA), and Strategies Analysis (StrA). The analysis synthesizes descriptions of M&T work drawn from literature review, participant observation, and the (limited) field study. In cognitive engineering terms, this analysis characterizes M&T as a task:

- That searches a work domain subject to continuous change (as equipment wears, equipment is replaced, business priorities change, or the M&T task succeeds)
- Which can include representation-maintenance in order to (hopefully) support assessing energy performance and making corrective changes
- That can be achieved with a combination of M&T strategies appropriate to particular task goals and circumstances
- That takes place in a social work environment, and can be distributed among colleagues

The findings of this CWA, such as a WDA information taxonomy, ConTA information flows, and StrA knowledge states can be applied to design knowledge bases, interfaces, and algorithms for M&T work support tools. ConTA characterizes a risk in M&T that tool designs must consider: minimizing labor consumed to maintain energy performance models' representation of a business structure that *must* change as part of improving energy-efficiency. StrA distinguishes M&T strategies dependent on data and model representations, and strategies that can use expert recognition alone. While this analysis relies on analyst judgment and was not validated, it is a starting point for theoretic investigation of the challenges inherent to M&T work. The analysis supports a design case for minimizing and managing sources of ambiguity in M&T diagnostic support tools, particularly by maximizing the benefit of the simplest energy data and performance models necessary.

### 6.1.3. Model summary sheet & modified Recursive Estimates

The third contribution of this dissertation in Chapter 4 is two novel M&T work support tools developed explicitly to support un-met work needs observed in the field study and characterized through work analysis.

The first work support tool, the model summary sheet, communicates information necessary to assess energy performance regression models in terms of their intended purpose, empirical basis, and range of validity. None of the three M&T work support tools I observed presented this comprehensive information, which would otherwise be known only to the analyst who developed the model. This addresses a knowledge gap which impedes distributed social organization of M&T. The second work support tool, a Modified Recursive Estimates (RE) diagnosis aid, builds on the Model Summary Sheet by adapting a statistical strategy to suit the M&T task and work environment. This statistical strategy

- has not been applied to M&T before,
- requires no extra representation-maintaining work since it re-uses existing linear energy performance models from standard M&T comparative analysis strategies,
- is robust and flexible and can be deployed even in a low-tech paper format, and
- supports diagnostic search through purpose-related functional work domain structures that existing strategies do not.

This work describes how cognitive engineering principles informed modifying RE charts to better suit the M&T task, by reformulating charts from dimensionless empirical fluctuation processes to analogical forms. The modified RE charts reduce ambiguity in visually representing estimated structural changes in the energy-consuming system. This demonstrates an alternate approach to coping with weaknesses of multiple linear regression such as a) measurement error and endogenous variables causing unidentifiable model parameters, and b) covariance among parameter changes. Instead of substituting more complex algorithms to fix flaws, statistical weaknesses can be mitigated by a) facilitating comparison against more robust methods (CUSUM charts) b) judicious formatting (invariant multiple-time scaling, and mapping salience to informativeness), and c) summarizing relevant information needed to make inferences from display features (the Model Summary Sheet). The first two design features should enable practitioners to develop locally useful interpretation rules and recognition-based expertise, while

the last should support knowledge-based behavior in diagnosing causes of energy waste. The content and form of modified RE chart displays is consistent with the theoretic principles of Ecological Interface Design. But RE charts might not be practically effective if they induce more false alarms, confusion, or mistaken diagnoses, are slower to learn, or slower to interpret. The dissertation investigates these concerns through an experimental synthetic M&T task.

#### 6.1.4. Controlled M&T task experiment

The final contribution of this dissertation, in Chapter 5, is the first published controlled experimental evaluation of M&T performance. Since M&T is performed in business systems which are intractable for experimental control and difficult to verify ‘ground truth’, a synthetic M&T task compliments field study methods. This experimental evaluation found that

- College students in an energy management professional program could learn to interpret RE charts after a 30 minute lecture and three practice trials
- Participants indicated that they found RE charts less confusing and more useful in distinguishing types of energy performance changes
- RE charts were associated with a twelve percent increase in task time
- False alarm rates were relatively high (over 50% of participants’ responses), but were not significantly different when using RE charts
- Participants attempted diagnosis often (76% of responses), and significantly more often (85%) when using RE charts
- Both change size (CUSUM slope) and accumulated evidence (vertical displacement) cues are associated with detecting changes
- Did not find evidence that any property of experimental changes was associated with being more diagnosable
- Diagnosis performance did not significantly differ from chance with CUSUM charts alone
- When using RE charts, participants diagnosed correctly 41% of the time, a significant improvement from industry-standard CUSUM charts alone

Some experimental conclusions could be drawn. The known statistical flaws of RE charts, which were unaffected by the design modifications, did not reduce change detection performance. However, neither did the redesign of visual formatting improve false alarm rates. RE charts

improved diagnosis performance, but the experiment could not determine whether the diagnosis benefits were associated with particular types of change (e.g. to variable heating consumption), other change properties (e.g. size), or individual differences in tactical rules (e.g. using provided visual masking aids with RE charts).

The high false alarm and diagnosis attempt rates observed in the experiment were unexpected, and should be investigated further. The experiment provided little prior information for participants to calibrate their tendency to respond, no individual feedback to calibrate self-efficacy, and took place in a no-stakes classroom setting. These are opportunities for future work. This experimental framework can be applied to evaluate other M&T work support tools, and could be extended to induce more representative behavior and simulate the cues needed for other strategies that more expert participants might switch between. Incidental contributions and opportunities for future work are discussed next.

## 6.2 Incidental contributions

Two incidental contributions from this dissertation have been (and will be) published elsewhere. The first is a description of the socio-organizational context of Energy Management through an interview study from Energy Managers, Program Developers and policymakers' perspective (Hilliard et al., 2009). The second is a set of minor CWA methods and theoretic innovations, specifically developing a novel CWA Work Domain Analysis approach to decomposition when analyzing categories of systems; extending CWA Control Task Analysis to represent abductive representation-maintaining as a conjoined control task; critiquing theoretical challenges of CWA Strategies Analysis; and making minor methods extensions. These contributions are mainly of interest to CWA theoreticians and practitioners (Hilliard and Jamieson, in review).

## 6.3 Limitations and future work

While the first published M&T field study, controlled experiment, and work analysis, this dissertation offers many opportunities for future work.

### 6.3.1. Field studies of true expert M&T behavior

The single field study conducted for this dissertation is just one data point, and suffered from sampling biases. Research in residential domains compare across many studies (Abrahamse et

al., 2005). To compliment the analyst-assisted M&T novices observed in this field study, future work should describe more expert M&T diagnosis behavior. This should include observations of more situated strategies (like Condition Surveys), user-developed information systems that satisfy (individuals') work needs, and businesses with high levels of employee engagement.

### 6.3.2. Future experimental work

The experimental structure demonstrated in this dissertation could be replicated to corroborate the results, examine other M&T diagnosis aids, and correct apparatus weaknesses. If participants performed the task individually, experiments could reduce individual differences and calibrate response biases by training to a performance criterion. Participation one-at-a-time would also facilitate an interactive software M&T task apparatus. Developing an interactive M&T task would improve data collection, such as of information sampling and navigation behaviors. The M&T task in this dissertation was developed to balance tractability and external validity. Future experiments could answer different research questions, either by isolating and controlling task performance or by enabling more flexible naturalistic behavior (Table 26).

**Table 26 - Dimensions on which experimental work described in this dissertation could be extended**

	<b>Dissertation work</b>	<b>Future work</b>
<i>Participant Expertise</i>	Practiced	Expert
<i>Timing</i>	Retrospective	Real-time
<i>Data acquisition</i>	Mediated only	including Situated
<i>Phenomena timescale</i>	Fixed	Multi-scale
<i>Problem space</i>	Bounded	Unbounded
<i>Decision Criteria</i>	Balance of probabilities	Risk/reward (social)

One future experiment could increase control, data collection, and ease of analysis by adopting classical psychology methods. Hundreds of single (or overlapping) stimuli could be constructed to fully counterbalance combinations of change properties. By using many independent stimuli, this would enable Signal Detection Theory sensitivity and bias constructs to be used in analysis. Such an experiment could provide evidence to:

- isolate properties of changes that might affect detectability or diagnosibility (magnitude, persistence, underlying trend , as in Section 5.2.2)
- isolate propensity to attempt diagnosis (or guess). This could be achieved by forcing participants to attempt a diagnosis, or limiting their allowed diagnosis attempts
- examine chart shapes to infer post-hoc what shapes were perceived as detection/diagnosis cues
- systematically vary change size to determine critical lower/upper thresholds for detection or diagnosis. In the experiment described in this dissertation, doubling change size more than doubled detection odds, but diagnosis performance was not affected
- systematically vary emergent features from particular combinations of overlapping changes in the scenario set. For example, change types could be systematically permuted (e.g. opposing vs. reinforcing changes)

A second M&T experiment could use a more naturalistic task environment with physical energy-consuming apparatus (such as an on-campus convenience store with electric grill, refrigerators, lighting, and icemaker). Faults could be deliberately introduced by experimenters to provide experimental control. Such an experiment could investigate:

- longitudinal effects of RE chart diagnosis representation aids for experienced participants. After learning has taken place, what chart features do participants begin relying on?
- M&T as a continuous monitoring task (rather than retrospective). How much evidence of a fault must accumulate before participants detect it?
- whether false alarm behaviors moderate as participants better understand variability in the underlying energy-consuming system?
- switches to physical search strategies (such as condition surveys)?
- use of contextual information - how to study information fusion of pre-existing beliefs or knowledge of historical events/conditions?

### 6.3.3. CWA Future work

There are many opportunities for future development of Cognitive Work Analysis methods and notations. Strategies Analysis in particular has been little-developed (c.f. Hassall & Sanderson, 2014) and could be extended to:

- represent cognition both ‘in the head’ and ‘in the world’, in order to describe situated, environment-driven action in compatible terms
- reflect contemporary theories of human cognition, such as the ‘adaptive toolbox’ view of environment-driven strategies (Gigerenzer, 2008), similar to CWA treatment of the Recognition-Primed Decision Making model (Naikar, 2010)
- expand methods advice for applying Strategies Analysis to design of information support tools (Hilliard & Jamieson, 2014a)

#### 6.3.4. Extending RE charts for energy performance diagnosis

RE charts as modified in this dissertation could be improved based on field and experimental assessment. Some changes include:

- refining the algorithm used to reduce the salience of un-informative RE chart segments. Instead of a moving window (Section 4.3.7.2), a retrospective non-lagging indicator (such as segmentation) would more clearly indicate times when a variable stagnated.
- similarly, mapping RE chart salience to a more information-theoretic measure than the relative variation compared to training used in this dissertation.
- exploring perceptual or statistical mitigations to allow RE charts to remain useful when applied to energy models with many parameters (and therefore more charts).
- exploring if RE charts can be extended to estimate changes to non-linear pre-processing of driver variables. For example, changes in the temperature ‘break point’ in calculating Heating Degree Days have engineering significance to changing thermostat setpoint.

Other improvements to RE charts could be developed in the context of a more interactive, multi-strategy M&T information system.

#### 6.3.5. Multi-strategy “ecological” M&T information systems

Considering the principles of Ecological Interface Design with respect to M&T work support suggests three opportunities for future work in developing information systems for M&T. First, an ecological information system should minimize the need for maintaining representations by leveraging as much knowledge-in-the-world as possible. This means supporting mobile devices and maximizing the value of epistemic (knowledge-seeking) action (Kirlik, 2006).

Second, where representation is unavoidable (such as utility meter data), data structures should be designed to map higher-order work domain constraints. This can allow using work domain constraints to more effectively maintain data using higher-order actions (i.e. understanding part-whole relationships to reduce the need for individual item categorization).

Finally, an ecological M&T database structure should be able to be maintained and applied with multiple strategies to balance effectiveness and effort in different situations. For example, workers on-site examining equipment should be supported in updating knowledge using on-site strategies. Similarly, databases should be applicable using as many strategies as possible, including strategies for thought experiments and forecasting. Some aspects of a more ecological M&T information system to investigate are listed in Appendix B.2.

## 6.4 Conclusions

Diagnosing energy efficiency performance changes is difficult, and under-addressed in the practice literature. Experimental and field study evidence corroborated that Comparative Analysis strategies with standard CUSUM charts are not effective diagnosis aids, at least for non-experts. This is in part because the CUSUM chart is a highly integrated cue, combining overlapping effects of disturbances, energy consuming equipment, and activities with an insufficiently explicated model. Depending on the pace of change in a business and data collection quality, energy models can introduce more ambiguity into energy performance metrics than they reduce. Novices may not be very effective at discriminating changes in CUSUM charts, as evidenced by many false alarms in the experimental study and few decisive responses in the field study. This is consistent with low sensitivity leaving peoples' tendency to identify changes to be influenced mainly by their response bias. In an experiment, where the consequences of mis-identifying changes were low, liberal bias manifested as a high false alarm rate. In the field study, where workers' professional reputation was at stake, bias manifested as reluctance to engage colleagues without complete trust in data and model quality. This suggests an opportunity for tools to improve non-expert M&T practitioners' sensitivity and ability to diagnose.

Effective human behavior, including detecting, diagnosing, and planning correct actions is supported by correct mental models. Field study evidence found that all three non-analyst

practitioners had incorrect understanding of how energy consumption was modeled. This impeded creative thought experiments, and may have contributed to ineffective diagnosis of CUSUM chart changes. The approach to supporting learning and developing expertise demonstrated in this dissertation is to expose, not conceal, the models underlying performance metrics. Linear regression energy performance models are relatively simple, and if clearly communicated in non-technical terms may serve as an evidence-grounded framework for workers to develop expert mental models of a particular site's energy consumption behavior. The Model Summary Sheet and modified RE chart diagnosis aids demonstrate this approach, exposing the system structure formulated in energy models to diagnostic search. This compliments existing search of physical functionality through utility sub-metering, or physical form through surveying business equipment condition.

Finally, just as all options must be considered in addressing climate change, all approaches to M&T diagnosis should be considered. The field study and work analysis of this dissertation examined M&T at multiple timescales, in systems with large problem spaces, where actors had to weigh social risk and reward. However, the experimental study in this dissertation omitted these factors (Table 26), constraining strategy interactions. Experimental analysis found that which types of changes were correctly diagnosed were strongly associated with participants' tendency to choose different diagnoses. However, the experiment design provided participants with few cues to form initial beliefs about which diagnosis changes were more probable. Other strategies can help develop these beliefs: condition surveys, event association, or equipment inventories can provide starting hunches for diagnosis. There is a need for management systems and tools to help socially develop these beliefs in an evidence-based manner while minimizing confirmation biases.

Reviewing a classic list of M&T challenges (Fawkes, 1988), it is notable how many still present today, e.g.:

- Distinguishing controllable from uncontrollable energy consumption variability
- Proving negatives – “Convincing people that they have actually saved energy when they have actually used more, and spent more, can be difficult” (Fawkes, 1988, p. 312)
- Framing energy performance in terms that managers will accept responsibility for
- Over-spending on technology that does not meet information needs

Modified RE chart methods have potential to distinguish energy consumption variability, and help communicate energy performance to colleagues. They complement utility energy sub-metering by helping initiate diagnostic searches in the functional structure described by existing CUSUM energy models, extracting more value from sunk costs of data collection and model maintenance. Future work, both field, experimental, analytical, and implementation can be pursued to achieve these goals.

Recursive estimates charts are a modest M&T diagnosis support extension to CUSUM charts. They complement CUSUM charts, being less statistically robust, but more informative, while remaining reasonably easy to interpret. Supporting a wide range of M&T work will be key for the success and adoption of tools such as the free National Renewable Energy Laboratory's Energy DataBus (NREL, 2015), Natural Resources Canada's RETScreen (NRCan, 2009), or the open-source EmonCMS (Emoncms, 2015). I hope this dissertation motivates development in M&T methods, and to apply these conclusions to practical M&T tools.



## References

- Abrahamse, W., Steg, L., Vlek, C., & Rothengatter, T. (2005). A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology, 25*(3), 273–291.
- Aird, R. J. (1981). Using degree days to manage energy. *Energy Management*.
- Allcott, H., & Greenstone, M. (2012). Is There an Energy Efficiency Gap? *Journal of Economic Perspectives, 26*(1), 3–28. <http://doi.org/10.1257/jep.26.1.3>
- Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: the “oblique effect” in man and animals. *Psychological Bulletin, 78*(4), 266–278.
- Ashby, W. R. (1956). *An introduction to cybernetics*. London; [New York]: Methuen; Distributed by Harper & Row.
- ASHRAE Guideline Project Committee 14P. (2002). *Guideline 14-2002: Measurement of Energy and Demand Savings* (1st ed.). Atlanta, GA: American Society of Heating, Refrigerating, and Air-Conditioning Engineers Inc.
- BC Hydro. (2010, December 3). BC Hydro - Industrial Energy Manager Initiatives. Retrieved February 1, 2011, from <http://www.bchydro.com/powersmart/industrial/offers/plan/iem.html>
- Bennett, K. B., & Flach, J. (2011). *Display and interface design: subtle science, exact art*. Boca Raton, Fla.: CRC Press.
- Bisantz, A. M., & Burns, C. M. (2009). *Applications of Cognitive Work Analysis*. Boca Raton: CRC Press.
- Bobker, M. (2004). Knowledge Practice in a Sea of Information. In B. L. Capehart (Ed.), *Information technology for energy managers* (pp. 171–178). Lilburn, Ga.: Fairmont Press.
- BRESCU. (2001). *Energy Management Priorities - a self-assessment tool* (No. GPG306) (p. 24). Garston, UK: Building Research Energy Conservation Support Unit. Retrieved from <http://www.carbontrust.co.uk/Publications/publicationdetail.htm?productid=GPG306>
- Bröder, A. (2003). Decision making with the “adaptive toolbox”: Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(4), 611–625. <http://doi.org/10.1037/0278-7393.29.4.611>
- Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for Testing the Constancy of Regression Relationships over Time. *Journal of the Royal Statistical Society, (35)*, 149–192.

- Bunse, K., Vodicka, M., Schönsleben, P., Brühlhart, M., & Ernst, F. O. (2011). Integrating energy efficiency performance in production management – gap analysis between industrial needs and scientific literature. *Journal of Cleaner Production*, 19(6-7), 667–679. <http://doi.org/10.1016/j.jclepro.2010.11.011>
- Burns, C. M., & Hajdukiewicz, J. R. (2004). *Ecological Interface Design*. Boca Raton, Florida: CRC Press.
- CanMET Energy. (2003). *Pinch Analysis: For the Efficient Use of Energy, Water, and Hydrogen*. Ottawa, Canada: Natural Resources Canada, Office of Energy Efficiency. Retrieved from [http://www.nrcan.gc.ca/sites/www.nrcan.gc.ca/files/canmetenergy/pdf/fichier.php/codectec/En/2009-052/2009-052\\_PM-FAC\\_404-DEPLOI\\_e.pdf](http://www.nrcan.gc.ca/sites/www.nrcan.gc.ca/files/canmetenergy/pdf/fichier.php/codectec/En/2009-052/2009-052_PM-FAC_404-DEPLOI_e.pdf)
- Capehart, B. L., & Capehart, L. (Eds.). (2005). *Web based energy information and control systems: case studies and applications*. Lilburn GA; Boca Raton FL: Fairmont Press; Distributed by Taylor & Francis.
- Capehart, B. L., Turner, W. C., & Kennedy, W. J. (2008). *Guide to energy management* (Vol. 6th). Lilburn, GA: Fairmont Press.
- Carbon Trust. (2006). *How to monitor your energy use* (No. GIL157) (p. 5). London: Queen's Printer and Controller of HMSO.
- Carbon Trust. (2007). *Advanced metering for SMEs: Carbon and cost savings* (No. CTC713). The Carbon Trust. Retrieved from <http://www.carbontrust.co.uk/Publications/publicationdetail.htm?productid=CTC713>
- Carbon Trust. (2008). *Monitoring and targeting: Techniques to help organizations control and manage their energy use* (No. CTG008). The Carbon Trust. Retrieved from <http://www.carbontrust.com/resources/guides/energy-efficiency/monitoring-and-targeting/>
- Chi, M. (1997). Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. *Journal of the Learning Sciences*, 6(3), 271–315. [http://doi.org/10.1207/s15327809jls0603\\_1](http://doi.org/10.1207/s15327809jls0603_1)
- CIPEC. (2010). *Energy Savings Toolbox - an Energy Audit Manual and Tool*. Canadian Industry Program for Energy Conservation. Retrieved from <http://www.nrcan.gc.ca/energy/efficiency/industry/cipec/5161>
- Cmar, G., & Gnerre, W. (2005). Defining the Next-Generation Enterprise Energy Management System. In B. L. Capehart & L. Capehart (Eds.), *Web based energy information and control systems: case studies and applications* (pp. 403–434). Lilburn GA; Boca Raton FL: Fairmont Press; Distributed by Taylor & Francis.
- Dalal, N. P., & Kasper, G. M. (1994). The design of joint cognitive systems: the effect of cognitive coupling on performance. *International Journal of Human-Computer Studies*, 40(4), 677–702. <http://doi.org/10.1006/ijhc.1994.1031>

- Dutton, J. M., & Starbuck, W. H. (1971). Finding Charlie's Run-Time Estimator. In *Computer Simulation of Human Behavior* (pp. 218–242). New York: John Wiley and Sons.
- Efficiency New Brunswick. (2010). *Energy Management Information Systems Planning Manual and Tool*. Natural Resources Canada. Retrieved from <http://www.nrcan.gc.ca/energy/efficiency/industry/cipec/5161>
- Efficiency Valuation Organization. (2012). *International Performance Measurement & Verification Protocol: Concepts and Options for Determining Energy and Water Savings* (7th ed., Vol. I). Efficiency Valuation Organization. Retrieved from [www.ipmvp.org](http://www.ipmvp.org)
- Emoncms. (2015). Emoncms. Retrieved July 21, 2015, from <http://emoncms.org/>
- Ericsson, K., & Simon, H. A. (1992). *Protocol analysis: verbal reports as data* (Rev. ed.). Cambridge Mass.: MIT Press.
- Fawkes, S. D. (1986). A comparison of British and Japanese industrial energy management. *R&D Management*, 16(4), 309–316. <http://doi.org/10.1111/j.1467-9310.1986.tb01177.x>
- Fawkes, S. D. (1988). Monitoring and targeting in buildings: Principles, programs and processors. *Applied Energy*, 29(4), 307–317. [http://doi.org/10.1016/0306-2619\(88\)90041-4](http://doi.org/10.1016/0306-2619(88)90041-4)
- Flemming, S. A. C., Hilliard, A., & Jamieson, G. A. (2008). The Need for Human Factors in the Sustainability Domain. In *Human Factors and Ergonomics Society 52nd Annual Meeting* (Vol. 52, pp. 748–752). New York, NY: HFES.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.
- Gigerenzer, G. (2001). *Simple heuristics that make us smart* (2. printing.). New York: Oxford University Press.
- Gigerenzer, G. (2008). *Rationality for mortals: how people cope with uncertainty*. New York; Oxford: Oxford University Press.
- Gotel, D. G., & Hale, D. K. (1989). *The application of monitoring & targeting to energy management*. London: HMSO.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Harris, P. (1989). *Energy monitoring and target setting using CUSUM*. Cheriton Technology Publications.
- Hassall, M. E., & Sanderson, P. M. (2014). A formative approach to the strategies analysis phase of cognitive work analysis. *Theoretical Issues in Ergonomics Science*, 15(3), 215–261. <http://doi.org/10.1080/1463922X.2012.725781>

- Henze, G. P. (2001). Building Energy Management as Continuous Quality Control Process. *Journal of Architectural Engineering*, 7(4), 97. [http://doi.org/10.1061/\(ASCE\)1076-0431\(2001\)7:4\(97\)](http://doi.org/10.1061/(ASCE)1076-0431(2001)7:4(97))
- Hilliard, A., & Jamieson, G. A. (2013). Recursive Estimates as an Extension to CUSUM-based Energy Monitoring & Targeting. In *Proceedings of the 2013 ACEEE Summer Study on Energy Efficiency in Industry* (pp. 4–1.4–13). Niagara Falls, NY: ACEEE. Retrieved from [http://aceee.org/files/proceedings/2013/data/papers/4\\_094.pdf](http://aceee.org/files/proceedings/2013/data/papers/4_094.pdf)
- Hilliard, A., & Jamieson, G. A. (2014a). A Strategy-Based Ecological(?) Display for Time-Series Structural Change Diagnosis. In *Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 353–358). San Diego, CA: IEEE.
- Hilliard, A., & Jamieson, G. A. (2014b). Monitoring & Targeting Energy in Practice: A Field Study. In *Proceedings of the 2014 ECEEE Summer Study in Industry* (pp. 591–601). Arnhem, NL: European Council for an Energy Efficient Economy. Retrieved from <http://www.eceee.org/library>
- Hilliard, A., Jamieson, G. A., & Jorjani, D. (2014). Communicating a Model-Based Energy Performance Indicator. *Ergonomics in Design: The Quarterly of Human Factors Applications*, 22(4), 21–29. <http://doi.org/10.1177/1064804614550861>
- Hilliard, A., Jamieson, G. A., & White, A. (2009). *Energy Management in Large Enterprises: A Field Study* (Technical Report No. CEL09-01). Toronto, Canada: Cognitive Engineering Laboratory. Retrieved from <http://cel.mie.utoronto.ca>
- Hooke, J. H., Landry, B. J., & Hart, D. (2004). *Energy management information systems - Achieving Improved Energy Efficiency: A handbook for managers, engineers and operational staff*. Natural Resources Canada, Office of Energy Efficiency. Retrieved from <http://publications.gc.ca/pub?id=290082&sl=0>
- Hutchins, E. (1996). *Cognition in the wild* (2nd printing.). Cambridge, MA: MIT Press.
- International Energy Agency. (2012). IEA Sankey Diagram. Retrieved August 8, 2015, from <http://www.iea.org/sankey/#?c=World&s=Final%20consumption>
- International Energy Agency. (2014). *Energy efficiency market report 2014: market trends and medium-term prospects*. Retrieved from <http://alltitles.ebrary.com/Doc?id=10961845>
- International Finance Corporation. (2011). Resource Efficiency in the Ferrous Foundry Industry in Russia: Benchmarking Study. IFC Advisory Services in Russia. Retrieved from <http://www.ifc.org/wps/wcm/connect/e5805e804bbee2b38ae8ef1be6561834/PublicationRussiaFoundry2011en.pdf>
- IPCC Core Writing Team. (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (p. 151). Geneva, Switzerland: IPCC. Retrieved from <http://ipcc.ch/report/ar5/syr/>

- ISO Technical Committee 242. (2011). ISO 50001:2011 - Energy Management. International Standards Organization. Retrieved from <http://www.iso.org/iso/iso50001>
- Jaffe, A. B., & Stavins, R. N. (1994). The energy paradox and the diffusion of conservation technology. *Resource and Energy Economics*, *16*(2), 91–122. [http://doi.org/10.1016/0928-7655\(94\)90001-9](http://doi.org/10.1016/0928-7655(94)90001-9)
- Jamieson, G. A., & Vicente, K. J. (2005). Designing effective human-automation-plant interfaces: A control-theoretic perspective. *Human Factors*, *47*(1), 12–34.
- Kempton, W. (1986). Two Theories of Home Heat Control. *Cognitive Science*, *10*(1), 75–90. [http://doi.org/10.1207/s15516709cog1001\\_3](http://doi.org/10.1207/s15516709cog1001_3)
- Kempton, W., Feuermann, D., & Mcgarity, A. (1992). “I always turn it on super”: user decisions about when and how to operate room air conditioners. *Energy and Buildings*, *18*(3-4), 177–191. [http://doi.org/10.1016/0378-7788\(92\)90012-6](http://doi.org/10.1016/0378-7788(92)90012-6)
- Kempton, W., & Montgomery, L. (1982). Folk quantification of energy. *Energy*, *7*(10), 817–827.
- Khodadadi, A., & Asgharian, M. (2008). *Change-point Problem and Regression: An Annotated Bibliography* (Working Paper No. 44) (p. 236). The Berkely Electronic Press. Retrieved from <http://biostats.bepress.com/cobra/art44>
- Kilgore, R., St-Cyr, O., & Jamieson, G. A. (2008). From Work Domains to Worker Competencies: A Five-Phase CWA for Air Traffic Control. In *Applications of Cognitive Work Analysis* (pp. 15–47). Boca Raton, FL: CRC Press.
- Kirlik, A. (1995). Requirements for Psychological Models to Support Design: Toward Ecological Task Analysis. In *Global perspectives on the ecology of human-machine systems* (Vol. 1, pp. 68–120). Hillsdale N.J.: L. Erlbaum Associates.
- Kirlik, A. (2006). Abstracting Situated Action: Implications for Cognitive Modeling and Interface Design. In A. Kirlik (Ed.), *Adaptive perspectives on human-technology interaction methods and models for cognitive engineering and human-computer interaction* (pp. 212–229). Oxford;; New York: Oxford University Press.
- Kissock, J. K., & Eger, C. (2008). Measuring industrial energy savings. *Applied Energy*, *85*(5), 347–361. <http://doi.org/10.1016/j.apenergy.2007.06.020>
- Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, & R. Calderwood (Eds.), *Decision Making in Action: Models and Methods* (pp. 138–147). Norwood, NJ: Ablex.
- Krämer, W., Ploberger, W., & Alt, R. (1988). Testing for structural change in dynamic models. *Econometrica: Journal of the Econometric Society*, 1355–1369.
- Krippendorff, K. (2004). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, *30*(3), 411–433. <http://doi.org/10.1093/hcr/30.3.411>

- Kuan, C.-M., & Chen, M.-Y. (1994). Implementing the fluctuation and moving-estimates tests in dynamic econometric models. *Economics Letters*, *44*, 235.
- Lau, N., & Jamieson, G. A. (2006). Numerical Models in Representation Design: Computing Seawater Properties in an Ecological Interface. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 245–249).
- Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.
- Lehrer, D., & Vasudev, J. (2010). Visualizing information to improve building performance: a study of expert users. Presented at the ACEEE Summer Study on Buildings, Pacific Grove, California, USA. Retrieved from <http://www.escholarship.org/uc/item/4n08r2q2>
- Lintern, G. (2009). The Foundations and Pragmatics of Cognitive Work Analysis. Retrieved from <http://www.cognitivesystemsdesign.net>
- Lofland, J., Snow, D. A., Anderson, L., & Lofland, L. H. (2006). Logging Data. In *Analyzing social settings: a guide to qualitative observation and analysis* (4th ed., pp. 81–117). Belmont CA: Wadsworth/Thomson Learning.
- Lutzenhiser, L. (1993). Social and Behavioral Aspects of Energy use. *Annual Review of Energy and the Environment*, *18*(1), 247–289.
- Lyle, O. (1947). *The Efficient Use of Steam*. London: HMSO.
- Macmillan, N. A. (1991). *Detection theory: a user's guide*. Cambridge [England]; New York: Cambridge University Press.
- Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review*, *118*(3), 393–437. <http://doi.org/10.1037/a0024143>
- Maynard, H. B., Stegemerten, G. J., & Schwab, J. L. (1948). *Methods-time measurement*. New York, NY, US: McGraw-Hill.
- McIlroy, R. C., & Stanton, N. A. (2015). Ecological Interface Design Two Decades On: Whatever Happened to the SRK Taxonomy? *IEEE Transactions on Human-Machine Systems*, *45*(2), 145–163. <http://doi.org/10.1109/THMS.2014.2369372>
- McKay, K. N. (1987). *Conceptual Framework for Job Shop Scheduling* (M.A.Sc.). University of Waterloo, Waterloo, ON, Canada.
- McKay, K. N. (1992). *Production Planning and Scheduling: A Model for Manufacturing Decisions Requiring Judgement* (Ph.D.). University of Waterloo, Waterloo, ON, Canada.
- Meier, A., Aragon, C., Hurwitz, B., Peffer, T., & Pritoni, M. (2010). How People Actually Use Thermostats (Vol. 2, pp. 193–206). Presented at the 2010 ACEEE Summer Study on Energy Efficiency in Buildings, Pacific Grove, CA, USA: American Council for an

- Energy Efficient Economy. Retrieved from <http://aceee.org/files/proceedings/2010/data/papers/1963.pdf>
- Meier, A., Aragon, C., Peffer, T., Perry, D., & Pritoni, M. (2011). Usability of residential thermostats: Preliminary investigations. *Building and Environment*, *46*(10), 1891–1898. <http://doi.org/10.1016/j.buildenv.2011.03.009>
- Mills, E. (2011). Building commissioning: a golden opportunity for reducing energy costs and greenhouse gas emissions in the United States. *Energy Efficiency*, *4*(2), 145–173. <http://doi.org/10.1007/s12053-011-9116-8>
- Moore, D. A. (2005). Sustaining Performance Improvements in Energy Intensive Industries. In *Proceedings of the Twenty-Seventh Industrial Energy Technology Conference* (pp. ESL–IE–05–05–31). New Orleans, LA.
- Moray, N. (1994). Ergonomics and the Global Problems of the 21st Century.
- Mumaw, R. J., Roth, E. M., Vicente, K. J., & Burns, C. M. (2000). There Is More to Monitoring a Nuclear Power Plant than Meets the Eye. *Human Factors*, *42*(1), 36–55.
- Naikar, N. (2010). A Comparison of the Decision Ladder Template and the Recognition-Primed Decision Model (DSTO-TR-2397). Air Operations Division, Defence Science and Technology Organisation. Retrieved from <http://www.dsto.defence.gov.au/attachments/A%20comparison%20of%20the%20decision%20ladder%20template%20and%20the%20recognition-primed%20decision%20model.pdf>
- Naikar, N., Hopcroft, R., & Moylan, A. (2005). *Work domain analysis theoretical concepts and methodology* (No. DSTO-TR-1665). Defence Science and Technology Organisation (Australia). Air Operations Division. Retrieved from <http://www.dtic.mil/dtic/tr/fulltext/u2/a449707.pdf>
- Naikar, N., Moylan, A., & Pearce, B. (2006). Analysing activity in complex systems with cognitive work analysis: concepts, guidelines and case study for control task analysis. *Theoretical Issues in Ergonomics Science*, *7*(4), 371–394. <http://doi.org/10.1080/14639220500098821>
- NRCan. (2009, January 10). RETScreen International. Retrieved from <http://www.retscreen.net>
- NREL. (2015). Energy Analysis - The Energy DataBus. Retrieved July 21, 2015, from <http://www.nrel.gov/analysis/databus/>
- Nyssen, A. S., & Javaux, D. (1996). Analysis of synchronization constraints and associated errors in collective work environments. *Ergonomics*, *39*(10), 1249–1264.
- Oehlert, G. W. (2000). *A first course in design and analysis of experiments*. New York: W.H. Freeman.

- Owens, C. (2013). *Multifamily Energy Auditor Job/Task Analysis and Report* (Subcontract Report No. NREL/SR - 7A40 - 60 447) (p. 94). Golden, CO: U.S. Department of Energy. Retrieved from <http://www.nrel.gov/docs/fy14osti/60447.pdf>
- Pacala, S. (2004). Stabilization Wedges: Solving the Climate Problem for the Next 50 Years with Current Technologies. *Science*, 305(5686), 968–972. <http://doi.org/10.1126/science.1100103>
- Page, E. S. (1961). Cumulative Sum Charts. *Technometrics*, 3(1), 1–9.
- Payne, J., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, New York, NY: Cambridge University Press.
- Perron, P. (2006). Dealing with Structural Breaks. In *Palgrave handbook of econometrics. 1, Econometric theory*. (pp. 278–353). Basingstoke: Palgrave Macmillan.
- Ploberger, W., Krämer, W., & Kontrus, K. (1989). A new test for structural stability in the linear regression model. *Journal of Econometrics*, 40(2), 307–318. [http://doi.org/10.1016/0304-4076\(89\)90087-0](http://doi.org/10.1016/0304-4076(89)90087-0)
- Rasmussen, J. (1974). *The human data processor as a system component: Bits and pieces of a model* (Technical Report No. Risø-M-1722) (p. 51). Roskilde, Denmark: Danish Atomic Energy Commission Risø.
- Rasmussen, J. (1979). *On the structure of knowledge - A morphology of mental models in a man-machine system context* (No. Risø-M-2192). Roskilde, Denmark: Risø National Laboratory.
- Rasmussen, J. (1986). *Information processing and human-machine interaction: an approach to cognitive engineering*. New York: North-Holland.
- Rasmussen, J., & Goodstein, L. P. (1987). Decision support in supervisory control of high-risk industrial systems. *Automatica*, 23, 663–671.
- Rasmussen, J., & Jensen, A. (1973). *A study of mental procedures in electronic trouble shooting* (No. Risø-M-1582). Roskilde, Denmark: Risø National Laboratory Electronics Department.
- Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering*. New York: Wiley.
- Rasmussen, J., Pejtersen, A. M., & Schmidt, K. (1990). *Taxonomy for cognitive work analysis*. Roskilde, Denmark: Risø National Laboratory.
- Rasmussen, J., & Vicente, K. J. (1989). Coping with human errors through system design: Implications for ecological interface design. *International Journal of Man-Machine Studies*, 31, 517–534.

- Reising, D. V. C., & Sanderson, P. M. (2002). Work domain analysis and sensors II: Pasteurizer II case study. *International Journal of Human Computer Studies*, 56(6), 597–637.
- R Project Team. (n.d.). The R Project for Statistical Computing. Retrieved September 16, 2014, from <http://www.r-project.org/>
- Russell, C. H. (2009). World-Class Energy Assessments. In A. Thumann, W. J. Younger, & T. Niehus, *Handbook of Energy Audits* (Eighth Edition). Linburn, GA: The Fairmont Press.
- Sandberg, P., & Söderström, M. (2003). Industrial energy efficiency: the need for investment decision support from a manager perspective. *Energy Policy*, 31(15), 1623–1634. [http://doi.org/10.1016/S0301-4215\(02\)00228-8](http://doi.org/10.1016/S0301-4215(02)00228-8)
- Sanderson, P. M., & Fisher, C. (1994). Exploratory sequential data analysis: foundations. *Human-Computer Interaction*, 9(3-4), 251–317.
- Schleich, J., & Gruber, E. (2008). Beyond case studies: Barriers to energy efficiency in commerce and the services sector. *Energy Economics*, 30(2), 449–464. <http://doi.org/10.1016/j.eneco.2006.08.004>
- Sheridan, T. B. (2006). Supervisory Control. In *Handbook of human factors and ergonomics* (Vol. 3rd, pp. 1025–1052). Hoboken, NJ: Wiley.
- Shipley, A. M., & Elliott, R. N. (2006). *Ripe for the Picking: Have we exhausted the low-hanging fruit in the Industrial Sector?* (No. IE061) (p. 28). American Council for an Energy-Efficient Economy.
- Sorrell, S., Mallett, A., & Nye, S. (2011). *Barriers to Industrial Energy Efficiency: A Literature Review* (No. V.11-87139) (p. 83). Vienna: United Nations Industrial Development Organization. Retrieved from [https://www.unido.org/fileadmin/user\\_media/Services/Research\\_and\\_Statistics/WP102011\\_Ebook.pdf](https://www.unido.org/fileadmin/user_media/Services/Research_and_Statistics/WP102011_Ebook.pdf)
- Stern, N. (2007). The Stern Review Report on the Economics of Climate Change. Retrieved from [http://www.hm-treasury.gov.uk/independent\\_reviews/stern\\_review\\_economics\\_climate\\_change/stern\\_review\\_report.cfm](http://www.hm-treasury.gov.uk/independent_reviews/stern_review_economics_climate_change/stern_review_report.cfm)
- Stuart, G., Fleming, P., Ferreira, V., & Harris, P. (2007). Rapid analysis of time series data to identify changes in electricity consumption patterns in UK secondary schools. *Building and Environment*, 42(4), 1568–1580. <http://doi.org/10.1016/j.buildenv.2006.01.004>
- Tanaka, K., Watanabe, H., & Endou, A. (2010). *Enerize E3 Factory Energy Management System* (No. 53.1) (p. 4). Japan. Retrieved from <http://cdn2.us.yokogawa.com/rd-te-r05301-005.pdf>
- Technological Economics Research Unit. (1979). *A study of the feasibility of energy costing and energy accounting models for management* (No. 15). University of Stirling.

- Therkelsen, P., McKane, A., Sabouni, R., Evans, T., & Scheihing, P. (2013). Assessing the Costs and Benefits of the Superior Energy Performance Program (Vol. 5, pp. 1–13). Presented at the 2013 ACEEE Summer Study on Energy Efficiency in Industry, Niagara Falls, NY: ACEEE. Retrieved from [http://eetd.lbl.gov/sites/all/files/aceee\\_sep\\_paper.pdf](http://eetd.lbl.gov/sites/all/files/aceee_sep_paper.pdf)
- Thollander, P., & Ottosson, M. (2010). Energy management practices in Swedish energy-intensive industries. *Journal of Cleaner Production*, *18*(12), 1125–1133. <http://doi.org/10.1016/j.jclepro.2010.04.011>
- Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. [http://doi.org/10.1016/0010-0285\(73\)90033-9](http://doi.org/10.1016/0010-0285(73)90033-9)
- U.S. Environmental Protection Agency. (2014, June 9). Draft Climate Controls Specification Version 1.0. Retrieved June 9, 2014, from [https://www.energystar.gov/products/specs/climate\\_controls\\_specification\\_version\\_1\\_0\\_pd](https://www.energystar.gov/products/specs/climate_controls_specification_version_1_0_pd)
- Vicente, K. J. (1998). Human factors and global problems: A systems approach. *Systems Engineering*, *1*(1), 57–69.
- Vicente, K. J. (1999). *Cognitive work analysis: toward safe, productive, and healthy computer-based work*. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Vicente, K. J. (2000). Toward Jeffersonian research programmes in ergonomics science. *Theoretical Issues in Ergonomics Science*, *1*(2), 93–112.
- Vicente, K. J., & Rasmussen, J. (1990). The Ecology of Human-Machine Systems II: Mediating “Direct Perception” in Complex Work Domains. *Ecological Psychology*, *2*(3), 207–249. [http://doi.org/10.1207/s15326969eco0203\\_2](http://doi.org/10.1207/s15326969eco0203_2)
- Vicente, K. J., & Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. *IEEE Transactions on Systems, Man and Cybernetics*, *22*(4), 589–606.
- Weather Underground. (n.d.). Weather History for Toronto Pearson, Canada. Retrieved from <http://www.wunderground.com/history/airport/CYYZ/2015/1/5/DailyHistory.html>
- Woods, D. D. (1991). The cognitive engineering of problem representations. In G. R. S. Weir & J. L. Alty (Eds.), *Human-computer interaction and complex systems*. London: Academic Press.
- Xiao, Y. (1994). *Interacting with Complex Work Environments: A Field Study and a Planning Model* (Ph.D.). University of Toronto, Graduate Department of Mechanical and Industrial Engineering, Toronto, Canada.
- Young, P. C. (2011). *Recursive estimation and time-series analysis*. Berlin: Springer-Verlag.
- Zeileis, A. (2003). Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, *44*(1-2), 109–123. [http://doi.org/10.1016/S0167-9473\(03\)00030-6](http://doi.org/10.1016/S0167-9473(03)00030-6)

Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). strucchange: An R Package for Testing for Structural Change in Linear Regression Models. *Journal of Statistical Software*, 7(2), 1–38.

Zuboff, S. (1988). *In the age of the smart machine: the future of work and power*. New York: Basic Books.



## Appendix A Cognitive Work Analysis of Energy M&T Addenda

### A.1 Comparing M&T to other Human Factors domains

These brief summaries outline features that I found differentiated M&T from domains where Human Factors theory and methods have been developed.

#### A.1.1 Nuclear Power

Nuclear power plants are a quintessentially specific, causal, static system engineered to mitigate high risk. Much 1970s/80s Human Factors work was performed in the nuclear power domain. Though nuclear plants are engineered for reliability and repeatable control, HF studies have concluded that monitoring for changes in plant condition involves active search and distinguishing between many simultaneous signals (Mumaw et al., 2000). This is similar to causal coupling in M&T, where by conservation of mass/energy, utility meters capture signals from every use of equipment in a business. Energy meters aggregate signals from downstream consumption, which can obscure the relationship between energy and phenomena of interest. M&T domains are usually less engineered than nuclear power plants and may have fewer sensors. M&T domains particularly differ from nuclear power plants in that energy efficiency is low-cost, not safety critical, open to disturbances, and must adapt to flexible business structure.

#### A.1.2 Aviation

Aviation includes several very different problem-solving domains, for example military air combat, civilian piloting, and air traffic control. Aviation systems are large-scope, but wide and shallow. Each individual plane is engineered for relatively predictable behavior, and system complexity arises from interactions between hundreds of aircraft and the environment. Aviation systems are mixed causal-intentional: causal for where flight is possible, intentional in where flight is desired. Like nuclear domains, aviation is safety-critical and therefore can justify high-cost solutions. Finally, like nuclear, aviation is focused on maintaining operation of an as-built engineered system.

### A.1.3 Safety / Risk management

Risk management addresses drifting change in a wide range of sociotechnical systems. Risk management has similar objectives to M&T in that practitioners want to evaluate opaque, difficult-to-measure system behavior against a desired standard. Risk management also has similar social constraints to M&T. Risk managers need to persuade colleagues to maintain or change work practices, even if less effortful alternatives are tempting. Risk management is more difficult than M&T because it lacks the benefit of causal conservation of mass/energy constraints that can enable statistical M&T diagnosis strategies.

### A.1.4 Quality Control

Quality control is quite similar to energy management. Quality is managed in many business domains, controls both causal and intentional constraints, and requires both detection and diagnosis. Quality Control differs from energy management in that quality is usually defined as a fixed target, and the goal of quality control is to reduce variability around that target. Energy management, however, aims to improve energy cost efficiency, which may *introduce* variability into energy use. This variability reflects the energy market cost environment, e.g. by varying electricity consumption to take advantage of low night-time prices.

ISO 9001-type quality control processes recommend controlling actions through procedures. However, this may not work for M&T interpretation since energy consumption is not shielded from disturbances as production lines are. Energy consumption is coupled to almost everything that happens in (or to) a business, so interpretation cannot be certain or routine. Management systems that mandate procedural compliance and acting only on statistically certain data may not adapt well to energy management.

## A.2 Some sensitizing concepts for CWA of M&T

Ashby's four conditions for system control (Section 3.2.1) suggests three sensitizing concepts to consider in CWA. First, the need for communication and coordination increase with social distribution of system control, understanding (mental models), and ability to observe the system. Communication and coordination can be supported through work tools, or reduced through social choices of centralizing control and sensors or decentralizing goals and (mental) models.

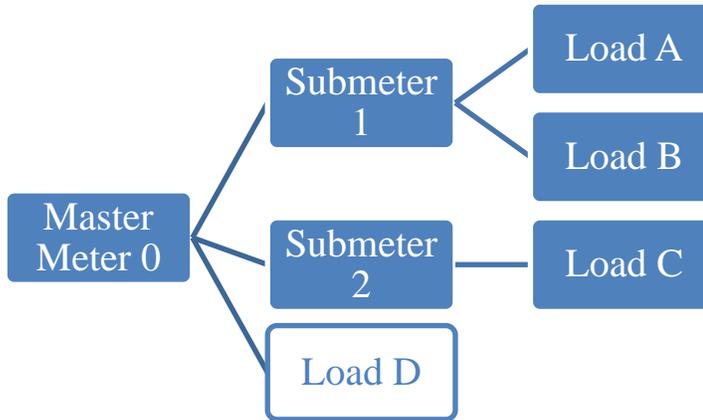
Second, a basic principle of systems theory states that no control system will perform better than its measuring channel (Ashby, 1956). Since utility energy meters measure effects of every energy-consuming process or activity in a business, they will contain what could be considered ‘noise’. A limiting factor in interpreting ‘noise’ into meaningful information is knowledge of context, or sufficiently complex (mental) models.

Finally, the granularity of control is a design option. While it may be possible to achieve a vision of immediate, dis-aggregated energy consumption feedback at nyquist-sufficient frequencies driving human (or automated) control, it is more common to control energy-consuming processes in ‘open loop’ and provide feedback on everyday consumption, aggregated by social group, in a common-sense approach with a clipboard and spreadsheet.

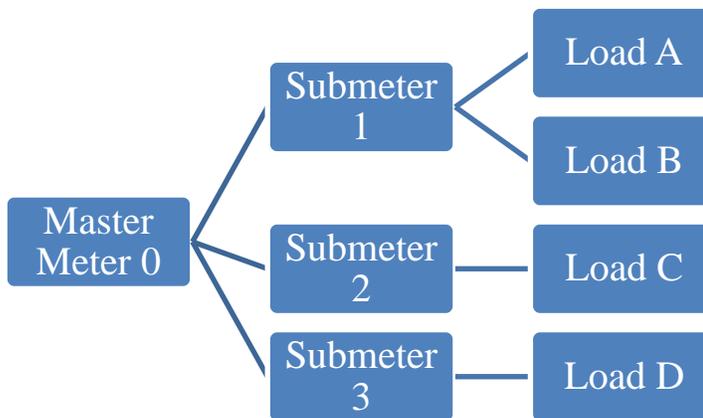
### A.3 Causal / Topographic structure in M&T

While work domain causal and topographic structures in are hard to analyze generically, there may be some common threads that could form a basis for information system design. These are:

- Physical Form & Function: Wire and pipe tracing
  - ◆ Utility supply networks usually have a ‘tree’ physical / functional topography. This means that when installing meter sensors, upstream (‘master’) meters capture all flow, some of which is redundantly captured by sub-meters.
- Abstract Function: Energy irreversibly flows from high to low ‘quality’, from supply to sinks (as embodied energy or heat dissipation)
  - ◆ Energy can be transformed several times, from primary (utility) supply, to secondary (e.g. compressed air) distribution, through tertiary (service delivery) to meet demand.



**Figure 44 – Fully- or Under-constrained metering system (where utility meter consistency cannot be checked)**



**Figure 45 - Over constrained metering system (can cross-check meter consistency aka. 'Unaccounted electricity')**

## A.4 Decision Ladder annotations

The Control Task Analysis (ConTA) of the Control Energy Costs and Cultivate Data & Models work functions (Section 3.4.3) includes a decision ladder (DL) diagram (Figure 14) briefly annotated with descriptions of the states of knowledge. These annotations are presented in more detail below in Table 27 and Table 28. They are phrased as questions whose answers (and the resulting state of knowledge) will be situation-specific. In doing the analysis, it seemed that strategy-specific questions were more helpful (e.g. in Section 3.5.4).

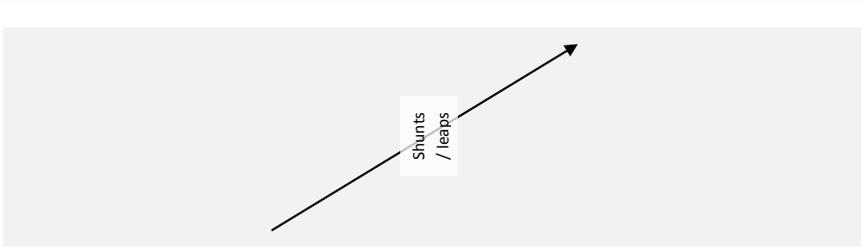
Situation Assessment		Annotated decision ladder for "Control Energy Costs" work function		Planning & Action	
<b>Goals</b>		To minimize net delivered energy cost to meet business requirements.			
<b>Ambiguity</b>	<p>How certain are equipment, process, or activity-related energy savings opportunities? How likely are alternative causes?</p> <p>Are the process principles addressed by the equipment consumption? Are there un-quantified benefits of current practice?</p> <p>What could be the reasons for variable consumption? What un-reported activities or equipment operation could have happened?</p>			<p>What could be the effects of misidentifying the situation on cost / benefits / meeting business requirements?</p> <p>Is it economic to improve equipment/system efficiency? Is it economic to match consumption and demand? How could billing rules make supply changes economic?</p> <p>How could business requirements be harmed if equipment / system was changed? What if situation develops on its own?</p>	<b>Goals</b>
<b>Aware of System State</b>	<p>&lt;&lt; <b>Overlaps with "Cultivate Data" function &gt;&gt;</b></p> <p>Why were activities done? How do activities affect the service demand? How much consumption was associated with that event/activity? How do the events/activities compare to the historic distribution of events/activities?</p> <p>How profitable was consumption across time / business areas?</p>			<p>What is a reasonable (Specific) Consumption? What is a reasonable net cost? What is a desirable amount/timing of service demand? What is a desirable amount/timing of consumption to meet service demand? How could processes be done more efficiently? What is a more desirable energy supply?</p>	<b>Desired State</b>
<b>Observations</b>	<p>&lt;&lt; <b>Overlaps with "Cultivate Data" function. &gt;&gt;</b></p> <p>Are there any leaks / drafts here? How much is the leak consuming? What activities happened today? What other disturbances happened?</p>			<p>What is the energy overconsumption to be corrected / savings opportunity to be pursued?</p> <p>What needs to be done to reduce energy waste? How can the wastage be corrected?</p>	<b>Required Task</b>
<b>Alerted</b>	<p>Does something look like energy waste that requires immediate attention? Did something interesting happen in the business? Has someone/something else indicated a problem?</p>			<p>What should be done to fix the energy-wasting equipment? What steps should be followed to behave energy-effectively?</p>	<b>Procedure</b>

Table 27 - Annotation of states of knowledge for "Control Energy Costs" work function shown in Figure 14.

Situation Assessment		Annotated decision ladder for "Cultivate Data & Models" work function		Planning & Action	
Goals		To maintain useful, trustworthy representations of business energy performance.		Goals	
<b>Ambiguity</b>	<p>How typical is the observed system state? What errors or omissions in observing or interpreting could have been made? How consistent is the timing / quantity / association of consumption data sources? How much consumption is unaccounted-for? What phenomena could have been obscured by sampling? What assumptions should be made in modeling? How valid are they? How do regularities in consumption structure compare to historic structure?</p>	<p>How could inconsistency, errors or omissions in knowledge of consumption, equipment structure, or business processes harm usefulness? Trust? Could changing energy performance models make them more useful? More trustworthy?</p>	<p>→</p>	<b>Consequences</b>	
<b>Aware of System State</b>	<p>How should metered consumption be adjusted, based on equipment / fuel quality / disturbances / meter condition?                      How does the 'type' of equipment function compare to its consumption? How does the equipment condition affect the consumption? What demands does the equipment function serve? How does the demand served by the equipment match its schedule? How does equipment nominal power draw and operation schedule indicate energy use? What is the 'Specific Consumption' per unit of demand?                      How does consumption compare to the historic distribution of demand? What is the 'Diversity factor'? What are regularities in consumption 'structure'?                      How do the billing rules produce delivered costs? What is the incremental cost of power/energy?</p>	<p>What is an achievable (Specific) Consumption? What is an achievable consumption profile?                      How should representations (data/models) be changed to be more useful or trustworthy?</p>	<p>→</p>	<b>Desired State</b>	
<b>Observations</b>	<p>What is the metered consumption? What condition is the meter in? What are the equipment operating parameters or adjustments? What temperature is the area / surroundings? What processes (production / material flow) happened?                      What are the pressure / temperature / area / flowrates? What service demand is there?                      What are the billing rules? What was the billed consumption?                      What equipment is here? What equipment do meters connect to? What function does equipment perform? What is the equipment's condition? What is the equipment's nominal power draw? What equipment is operating now, and on what schedule?</p>	<p>&lt;&lt; Not modeled &gt;&gt;                      Which recorded data points should be changed to be more useful or trustworthy?                      How should they be changed - made null or estimated?                      How can actions on domain expose or create desired information?</p>	<p>→</p>	<b>Required Task</b>	
<b>Alerted</b>	<p>Is there new data to collect? Has equipment arrangement changed? Have processes changed? Are better data/models needed for another activity?</p>	<p>&lt;&lt; Not modeled &gt;&gt;                      What steps should be taken to change recorded data?                      What steps should be taken to reveal information?</p>	<p>→</p>	<b>Procedure</b>	

Table 28 - Annotation of states of knowledge for "Cultivate Data & Models" work function shown in Figure 14.

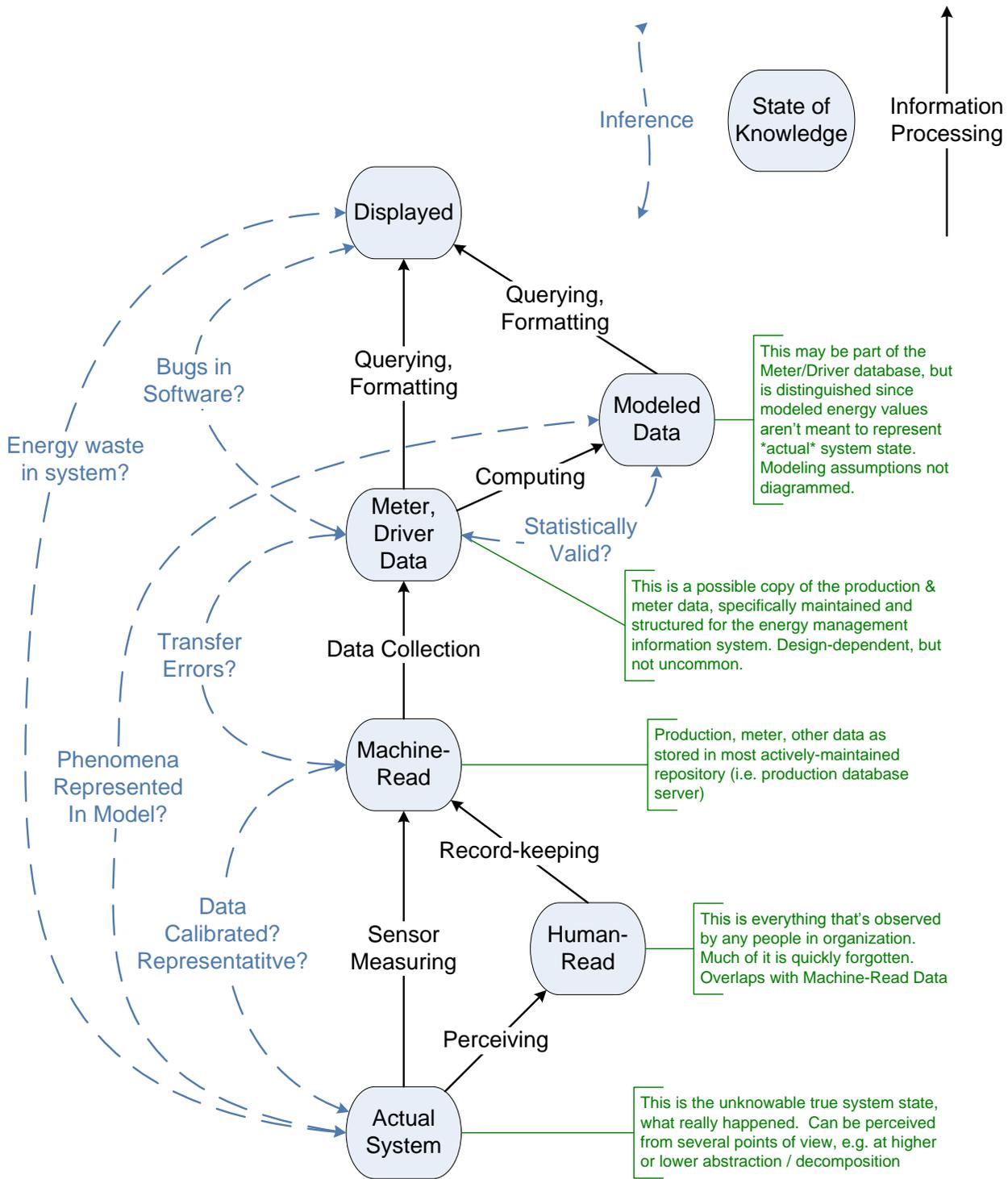
## A.5 M&T Data inference problems

The ‘Cultivating data and models’ work function was discussed in Section 3.4.3. The inferential aspects of this task, particularly relevant to Comparative Analysis strategies, are difficult to represent in DL notation. They are discussed here with reference to Figure 46 below.

In any record-keeping system, sensors, data processing, and representation can introduce distortions. Information processing steps diagrammed in Figure 46 describe overlap between several inference questions that were observed in the field study (Section 2.5.2):

- How much Energy Waste in System? – What does the data represented in an M&T software interface suggest about energy waste in the true system? Answering this question with certainty requires considering:
- Are Data Calibrated and Representative? – Do the energy meter and energy driver data consistently indicate the same energy-consuming phenomena? Has the system been reconfigured, record-keeping processes changed, or sensors mis-calibrated at any point?
- Are there any data Transfer Errors? – It is possible in some software designs for data to be missing because of technical errors. Low consumption may simply mean that a database connection was lost.
- How Statistically Valid are model-derived data? – As discussed in Section 2.3.3, statistical models transform data into representations of some system state (e.g. consumption at ‘historic performance level’). The models used to compute performance indicators may fail in certain circumstances.
- Are any Query or Formatting bugs distorting the displayed data? – This question is very design-dependent, but it is possible for there to be mistakes made in querying energy or performance indicator data.

Energy M&T software tools must either be infallible in processing data (which in the case of sensor quality or record-keeping may require an organizational solution) or support people inspecting intermediate steps of knowledge and assuring themselves of the integrity and trustworthiness of the data stream (Section 3.7.5).



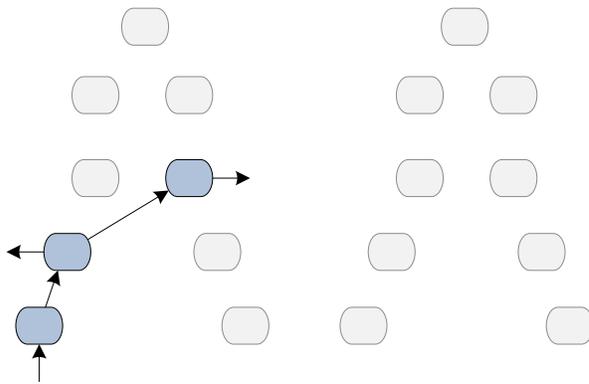
**Figure 46 - Some potential inferences (dashed lines) relevant to interpreting system state from energy data. Inconsistencies in information processing are what the work function of "Cultivating Data & Models" seeks to understand and minimize.**

## A.6 Instantiated Decision Ladders

Four examples below illustrate how the Decision Ladder notation of energy M&T functions in Figure 14 and the corresponding knowledge states in Appendix A.1 can capture several distinct tasks. All these tasks (except Energy Audit) are also considered as strategies in Section 3.5.

### A.6.1 “Automatic” Monitoring & Targeting

This sub-task is the automated calculation of achievable energy performance (Hooke et al., 2004), omitting model-building, interpretation, and corrective action.



**Figure 47 – Elements of Decision Ladder (Figure 14) active in "Automatic" Monitoring & Targeting subtask. Three states of knowledge are relevant: Alerted, Observations, and Desired State.**

The sequence of four knowledge-transforming processes (arrows in Figure 47) are:

- 1) At some time interval (minutes or days), automatic scripts (or a human operator) collect data from sensors or databases
- 2) These observations are compiled in an energy performance model database (which can be queried by others)
- 3) A pre-specified energy performance model is applied to transform the observations into a target energy consumption
- 4) Calculated desired energy consumption is compiled in a database and communicated to other actors

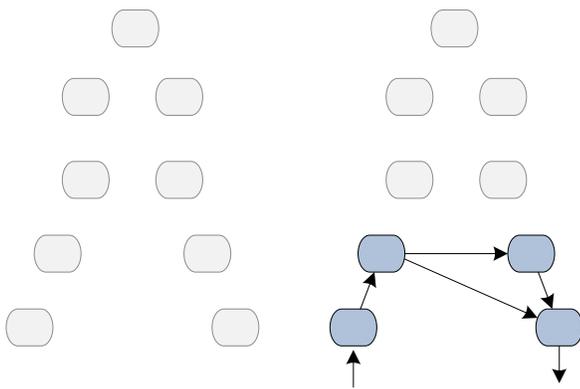
Some properties of this subtask are:

- Routine
- Described at length (Section 2.3) in literature

- Not a sufficient control task
- Does not consider ambiguity – just routine data processing
- Many task steps left incomplete.

### A.6.2 Condition survey

This subtask is the routine detection and repair of energy-wasting leaks, squeaks, or idling. It is described as part of Japanese energy management culture (Section 2.2.3).



**Figure 48 – Elements of Decision Ladder (Figure 14) active in “Condition survey” subtask. Relevant states of knowledge: Alerted, Observations, Task, Procedure.**

This subtask requires no models, data record, or assessment of system state. Six transitions between knowledge states are:

- 1) A worker is alerted by noticing something amiss
- 2) They observe the condition or operation of the equipment (e.g. lights on in an unoccupied space, leak)
- 3) They recognize or associate the condition with either a procedure (turn off the light) or a task (repair the leak)
- 4) If required, they plan a procedure (What needs to be shut off to fix the leak?)
- 5) They take action as soon as practicable

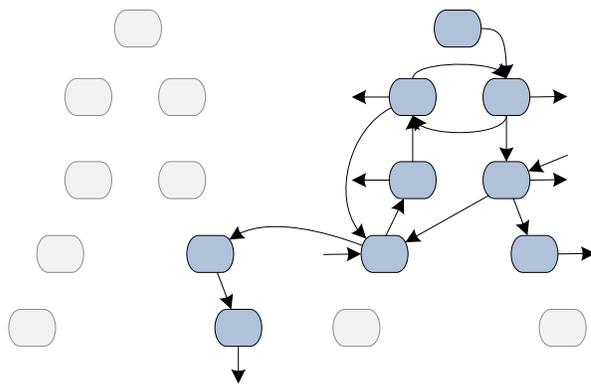
Some properties of this subtask are:

- Routine, does not require looking beyond observations to system state
- Requires practice and expertise to associate observations with repair tasks or procedures
- Requires mindfulness to be alerted by out-of-place conditions

- Can only control energy waste that is directly indicated by perceptible cues (ecological). Sensors (e.g. infrared cameras for heat energy, ultrasonic microphones for gas leaks) can extend cues humans can perceive.

### A.6.3 Energy Efficiency Audit

This subtask represents an engineering energy efficiency “audit”, typically performed by outside consultants. While this subtask does not actually fix problems, it develops recommendations (e.g. replace equipment) for management action. The subtask is illustrated in Figure 49 as a subset of the “Control Energy Costs” function, since it is oriented towards saving energy. Collected observations are used to develop specific plans, not stored in a database maintained for later use.



**Figure 49 – Elements of Decision Ladder (Figure 14) active in “Energy Audit” subtask**

Some transitions between knowledge states are:

- 1) The subtask can start with knowledge of desired state (what needs to happen for this business to achieve a certain efficiency), or with observations unrelated to a desired end state
- 2) Uncovering hidden observations may require acting on the system (e.g. test-running equipment, or installing temporary electric meters), shown as a leap to the ‘cultivate data’ goal
- 3) The task outputs are records of inferred system state, deliberations, desired state, and task specifics for (hopefully) later implementation.

As mentioned above, this subtask is shown distinct from “Cultivate Data & Models” to indicate that it is oriented towards planning energy-saving tasks. Representations are only developed for a point in time.



- Equipment condition survey
- Equipment inventory
- Consumption profile pattern-detection
- Comparative analysis
- Energy consumption & cost analysis
- External event association

Switches between strategies are an opportunity for information system support. Each strategy uses different observations or representations, which may be maintained or acted on using different strategies. Some interactions could include:

- Equipment Inventory <> Consumption Profile:
  - ◆ List of equipment and nameplate / measured current draw can be reconciled with profile signal processing, identifying step-changes in meter consumption profile.
  - ◆ Power step-change size can be monitored to try and deduce load / efficiency based problems.
  - ◆ Estimated run-time can be reconciled with deduced run time.
  - ◆ Individual equipment or Sub-metering can auto-deduce where it is attached by comparing its consumption profile to other meters, looking for time-domain correlation. This can be reconciled with the observed metering network.
- Equipment Inventory <> Condition Survey:
  - ◆ Equipment list and meter connectedness can be maintained during in-the-field observation.
  - ◆ Equipment list can be annotated with historic photographs to track condition degradation over time.
  - ◆ Video feeds from shop floor can be linked to navigation through equipment inventory
- Equipment Inventory <> Comparative Analysis
  - ◆ Changes in either should be marked and reconciled.
  - ◆ Equipment inventory can be compared with KPIs (model parameters) to compare across sites.
- Condition Survey <> Consumption & Cost Analysis:
  - ◆ Cost of consumption can be estimated on-site to help prioritize repairs (including conversion factors for compressed air, water leaks)

- Condition Survey <> Consumption Profile:
  - ◆ Observed (or induced) switching on/off equipment can be noted and used in meter signal processing.
  - ◆ Deduced signal processing equipment start/stop times can be inter-related to security camera or control data log timestamps.
- Comparative Analysis <> Consumption Profile:
  - ◆ Multi-level / timescale models can be created / reconciled. For example, modeled daily Baseload can be compared against 'true baseload'.
  - ◆ Autocorrelated historical 24h profiles can be augmented / reconciled with variable-driven models to allow 'zooming in'.
- External Event awareness <> All :
  - ◆ Known events can be annotated with time / area affected to allow blame / credit to be assigned.

## Appendix B Statistical Change Detection Discussion

### B.1 Test Statistics in CUSUM charts

#### B.1.1 Change detection

Test statistics exist for calculating the certainty of a system change having occurred, given a CUSUM chart (Page, 1961; Zeileis, 2003). However I did not observe or find any literature describing M&T practitioners using them, outside of Measurement and Verification (M&V) calculations (ASHRAE Guideline Project Committee 14P, 2002). This could be due to:

- Lack of knowledge among non-academic practitioners.
- Difficulty of explaining test statistic calculation and interpretation to end-users.
- Difficulty of implementing test statistics into commodity office tools (e.g. spreadsheets)
- Inability to distinguish changes. The pace of change in most businesses means will introduce changes that are outside of the energy managers' control. If these changes cannot be statistically distinguished from 'unknown' changes, findings will be obvious and unhelpful. Incorporating or correcting for these changes in the model might be technically possible, but not time-effective.
- A lack of informativeness. Certainty and significance are not equivalent. Detecting very minute changes may not be necessary, since such small changes may have little financial impact. Likewise, detecting minute changes may be less informative than the exact magnitude of the change (as for enforcing financial contract terms).

Whatever the reason, this is evidence that M&T work has distinct needs from those of formal structural change detection.

### B.2 'Ecological' Information system features for M&T support

#### B.2.1 Support units compatible with physical sampling

An irony of automatic data collection for M&T is that the less workers have to go out and measure, the less qualitative data about business operations they incidentally observe, limiting their ability to interpret data. The units in which utility and driver data are represented are a design opportunity in an 'ecological' M&T information system. Effective units should support

physical on-site data sampling, and help develop expert recognition of 'normal' energy consumption patterns.

Future work might evaluate whether an M&T information system should by default present data in time-invariant instantaneous units, for example average power, energy per unit production, or production rate rather than energy and production totals. This was proposed by early M&T researchers (Gotel & Hale, 1989, p. 5) but in my experience has not been adopted by vendors.

Reasons to adopt time-invariant instantaneous units are:

- 1) They're compatible with physical sampling. In a factory, workers can walk out to the plant floor and directly sample cues related to specific power consumption or rate.
- 2) The units used to normalize power or productive rate can be compatible with existing work domain purposes. Production rate can be the primary objective of some factory staff
- 3) Invariant units help develop associative long-term memory with factors that influence energy use: system state or equipment operation patterns. Experts can develop 'lookup table' type mental models of equipment operating patterns (Dutton & Starbuck, 1971)
- 4) Mental models in terms of invariant units provide instant functional context for assessing or 'reality checking' data. Equipment has functional limits of power draw, specific energy consumption and production rate.
- 5) It is usually valid over timescales where Comparative Analysis is performed (greater than daily), and allows comparing irregularly-timed periods

Two caveats should be considered. First, averaging in time-invariant units assumes a comparable operating mode for the entire time period. Consideration should be given to reasoning about operating modes, e.g. a factory stopped or running. Second, specific energy variables (like kWh/ton) are extensive properties so cannot be summed or arithmetically averaged, so they are not suited to database storage or record-keeping operations. Offering easy presentation in time-invariant units should reduce the effort needed to switch between data-driven and condition survey strategies.

### B.2.2 Support contextual time reference frames not just calendar dates

A similar barrier to condition survey strategies, as well as profile pattern strategies, is supporting multiple temporal reference frames (Nyssen & Javaux, 1996). Rather than only calendar dates, energy consumption is associated with:

- Work shifts
- Production cycles
- Customer patterns, such as weekends, holidays, and special events
- Operating modes

An ecological M&T information system should support switching between temporal reference frames as workers require.

### B.2.3 Supports social engagement?

An ecological M&T information system should be social. For example, utility and driver data assessment could be supported by meta-data about colleagues' inspection and endorsement. This could incorporate automated agents performing context-free statistical data checking. Data quality endorsements could be granular by timescale, utility sub-meter area, date ranges, etc. This provides an opportunity to delegate data assessment to those most able to assess its face validity.

### B.2.4 Forecasting and thought experiments?

An ecological M&T information system could support model-making not just as a carefully curate model, but as disposable though experiment tools for question-answering or thought experiments. This would require easily assessed utility meter data quality, and a software tool that supports evaluating model assumptions. The result could allow models to be tailored to answer specific work questions as attempted in the field study(Hilliard & Jamieson, 2014b).

Wherever possible, a software tool should support perceptually comparing models. For example, an old M&V model with a steady over/underconsumption could be compared with a more recent operation-guiding model by rotating CUSUM chart axes.



## Appendix C Experiment I Statistical Output

### C.1 Experimental Design

#### C.1.1 Model Training Data

The weather data used for the training set was drawn from Environment Canada Toronto Pearson Airport records (Weather Underground, n.d.). The cut-in break point for the heating system was arbitrarily set at below 12C (with 1% introduced error). The consumption data also included a pre-computed periodic disturbance created by four overlapping sinusoids (period 216, 108, 72, and 31 days). These sinusoids proved a poor change obfuscation tool and were omitted in Experiment II. Below are summary statistics of the first year's data (starting January 2009) used to train the baseline model.

```
> describe(OutputFull[which(OutputFull$Timestamp <
as.Date("2010/01/01")),][,c("vars", "n", "mean", "sd", "median", "min", "max", "range", "se"
)]]
```

	vars	n	mean	sd	median	min	max	range	se
Timestamp*	1	365	NaN	NA	NA	Inf	-Inf	-Inf	NA
Temperature	2	365	8.03	10.16	8.70	-17.60	27.20	44.80	0.53
Precipitation	3	365	2.48	6.02	0.00	0.00	40.20	40.20	0.32
IsWeekday*	4	365	0.72	0.45	1.00	0.00	1.00	1.00	0.02
Disturbance	5	365	9.21	123.02	13.31	-220.83	220.67	441.50	6.44
HDD	6	365	6.40	7.67	3.30	0.00	29.60	29.60	0.40
HDDTrue	8	365	6.40	7.68	3.39	0.00	29.73	29.73	0.40
Consumption	10	365	9123.69	3004.35	8170.86	5408.13	17851.21	12443.08	157.25
Base	11	365	6001.00	307.35	5991.15	5104.06	7020.26	1916.20	16.09
HDDCoef	12	365	369.98	17.85	368.62	307.69	438.20	130.52	0.93
HDDBreak	14	365	12.02	0.12	12.02	11.66	12.47	0.81	0.01
WeekdayCoef	16	365	1.00	0.01	1.00	0.97	1.03	0.06	0.00
PrecipCoef	17	365	185.10	9.02	185.43	154.64	211.13	56.49	0.47
WeibullError	18	365	276.93	210.83	227.76	2.88	1158.29	1155.41	11.04

### C.2 Data Summaries

#### C.2.1 Performance Data, Aggregated by-Participant

Below is a summary of the Experiment I metrics aggregated at Participant-level. Measures ending with `_inbox` and `_likely` refer to the scoring rules discussed in Section 5.2.8.

```
> print(describe(PLevelFactors)[,c("vars", "n", "mean", "sd", "median", "trimmed",
"mad", "min", "max", "range", "skew")])
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
FirstBooklet*	1	36	0.50	0.50	0.50	0.50	0.74	0.00	1.00	1.00	0.00
Participant*	2	36	10.22	5.62	10.50	10.27	7.41	1.00	19.00	18.00	-0.10
Order*	3	36	2.56	1.18	3.00	2.57	1.48	1.00	4.00	3.00	-0.13
School*	4	36	1.56	0.50	2.00	1.57	0.00	1.00	2.00	1.00	-0.21
Time.min	5	36	31.33	12.60	29.00	29.80	8.90	13.00	68.00	55.00	1.28
TrueScenarios*	6	36	1.50	0.51	1.50	1.50	0.74	1.00	2.00	1.00	0.00
Interface*	7	36	1.50	0.51	1.50	1.50	0.74	1.00	2.00	1.00	0.00
Responses	8	36	17.61	5.01	16.50	17.60	5.19	9.00	27.00	18.00	0.12
TrueChanges	9	36	15.00	0.00	15.00	15.00	0.00	15.00	15.00	0.00	NaN
MarkedCause	10	36	12.86	5.56	12.00	12.93	4.45	0.00	23.00	23.00	-0.05
Change_inbox	11	36	6.81	1.95	6.00	6.70	1.48	3.00	12.00	9.00	0.51
Diag_inbox	12	36	5.06	2.24	5.00	5.20	1.48	0.00	10.00	10.00	-0.54
RightDiag_inbox	13	36	3.06	1.71	3.00	3.00	1.48	0.00	8.00	8.00	0.59
Change_likely	14	36	10.03	1.98	10.00	10.17	1.48	5.00	13.00	8.00	-0.68
Diag_likely	15	36	7.47	2.75	8.00	7.77	2.97	0.00	11.00	11.00	-0.92
RightDiag_likely	16	36	4.22	2.04	4.00	4.27	1.48	0.00	8.00	8.00	-0.19
H	17	36	10.03	1.98	10.00	10.17	1.48	5.00	13.00	8.00	-0.68
FA	18	36	7.58	4.27	6.50	7.37	3.71	1.00	18.00	17.00	0.48
M	19	36	4.97	1.98	5.00	4.83	1.48	2.00	10.00	8.00	0.68
AD	20	36	12.86	5.56	12.00	12.93	4.45	0.00	23.00	23.00	-0.05
HD	21	36	7.47	2.75	8.00	7.77	2.97	0.00	11.00	11.00	-0.92
RD	22	36	4.22	2.04	4.00	4.27	1.48	0.00	8.00	8.00	-0.19
WD	23	36	3.25	1.73	3.00	3.17	1.48	0.00	8.00	8.00	0.49
Hr	24	36	0.67	0.13	0.67	0.68	0.10	0.33	0.87	0.53	-0.68
FAr	25	36	0.40	0.14	0.39	0.40	0.14	0.11	0.78	0.67	0.27
ADr	26	36	0.73	0.23	0.78	0.76	0.20	0.00	1.00	1.00	-1.37
HDr	27	36	0.75	0.25	0.82	0.79	0.17	0.00	1.00	1.00	-1.39
RDr	28	35	0.54	0.21	0.57	0.56	0.21	0.00	0.83	0.83	-0.85
CRr	29	36	0.25	0.13	0.27	0.26	0.16	0.00	0.50	0.50	-0.19

## C.2.2 Performance Data, by-Participant and Interface

```
print(describeBy(PLevelFactors, PLevelFactors$Interface))
```

```
## group: C
```

##	vars	n	mean	sd	median	trimmed	mad	min	max	range
## FirstBooklet*	1	18	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf
## Participant*	2	18	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf
## Order*	3	18	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf
## School*	4	18	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf
## Time.min	5	18	29.39	13.64	27.00	28.00	9.64	13.00	68.00	55.00
## TrueScenarios*	6	18	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf
## Interface*	7	18	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf
## Responses	8	18	17.00	5.22	15.50	16.88	5.19	9.00	27.00	18.00
## TrueChanges	9	18	15.00	0.00	15.00	15.00	0.00	15.00	15.00	0.00
## MarkedCause	10	18	12.17	5.84	11.50	12.12	5.93	2.00	23.00	21.00
## Change_inbox	11	18	6.67	2.30	6.00	6.56	1.48	3.00	12.00	9.00
## Diag_inbox	12	18	4.67	2.61	4.50	4.62	2.22	0.00	10.00	10.00
## RightDiag_inbox	13	18	2.72	1.78	2.00	2.56	1.48	0.00	8.00	8.00
## Change_likely	14	18	9.67	2.35	10.00	9.75	1.48	5.00	13.00	8.00
## Diag_likely	15	18	6.83	2.96	6.00	6.94	3.71	1.00	11.00	10.00
## RightDiag_likely	16	18	3.56	2.01	3.50	3.50	2.22	0.00	8.00	8.00
## H	17	18	9.67	2.35	10.00	9.75	1.48	5.00	13.00	8.00

```

## FA          18 18  7.33  4.75  5.50   7.06 3.71  1.00 18.00 17.00
## M           19 18  5.33  2.35  5.00   5.25 1.48  2.00 10.00  8.00
## AD          20 18 12.17  5.84 11.50  12.12 5.93  2.00 23.00 21.00
## HD          21 18  6.83  2.96  6.00   6.94 3.71  1.00 11.00 10.00
## RD          22 18  3.56  2.01  3.50   3.50 2.22  0.00  8.00  8.00
## WD          23 18  3.28  1.96  3.00   3.12 2.22  1.00  8.00  7.00
## Hr          24 18  0.64  0.16  0.67   0.65 0.10  0.33  0.87  0.53
## FAr        25 18  0.40  0.17  0.38   0.39 0.17  0.11  0.78  0.67
## ADr        26 18  0.70  0.22  0.74   0.71 0.20  0.18  1.00  0.82
## HDr        27 18  0.71  0.25  0.75   0.72 0.21  0.14  1.00  0.86
## RDr        28 18  0.49  0.25  0.47   0.50 0.22  0.00  0.83  0.83
## CRr        29 18  0.22  0.14  0.22   0.22 0.13  0.00  0.46  0.46

## -----
## group: C+R
##          vars  n  mean   sd median trimmed  mad   min  max range
## FirstBooklet*    1 18   NaN   NA   NA     NaN   NA   Inf -Inf -Inf
## Participant*    2 18   NaN   NA   NA     NaN   NA   Inf -Inf -Inf
## Order*          3 18   NaN   NA   NA     NaN   NA   Inf -Inf -Inf
## School*         4 18   NaN   NA   NA     NaN   NA   Inf -Inf -Inf
## Time.min        5 18 33.28 11.52 30.50  32.06 7.41 19.00 67.00 48.00
## TrueScenarios*  6 18   NaN   NA   NA     NaN   NA   Inf -Inf -Inf
## Interface*       7 18   NaN   NA   NA     NaN   NA   Inf -Inf -Inf
## Responses       8 18 18.22  4.86 19.50  18.31 5.93 10.00 25.00 15.00
## TrueChanges     9 18 15.00  0.00 15.00  15.00 0.00 15.00 15.00  0.00
## MarkedCause    10 18 13.56  5.33 12.50  13.81 3.71  0.00 23.00 23.00
## Change_inbox   11 18  6.94  1.59  6.50   6.81 2.22  5.00 11.00  6.00
## Diag_inbox     12 18  5.44  1.79  5.00   5.62 0.74  0.00  8.00  8.00
## RightDiag_inbox 13 18  3.39  1.61  3.00   3.38 1.48  0.00  7.00  7.00
## Change_likely  14 18 10.39  1.50 10.50  10.44 0.74  7.00 13.00  6.00
## Diag_likely    15 18  8.11  2.45  8.50   8.44 1.48  0.00 11.00 11.00
## RightDiag_likely 16 18  4.89  1.91  5.00   5.00 1.48  0.00  8.00  8.00
## H              17 18 10.39  1.50 10.50  10.44 0.74  7.00 13.00  6.00
## FA             18 18  7.83  3.85  8.00   7.75 5.93  3.00 14.00 11.00
## M              19 18  4.61  1.50  4.50   4.56 0.74  2.00  8.00  6.00
## AD             20 18 13.56  5.33 12.50  13.81 3.71  0.00 23.00 23.00
## HD             21 18  8.11  2.45  8.50   8.44 1.48  0.00 11.00 11.00
## RD             22 18  4.89  1.91  5.00   5.00 1.48  0.00  8.00  8.00
## WD             23 18  3.22  1.52  3.00   3.25 1.48  0.00  6.00  6.00
## Hr             24 18  0.69  0.10  0.70   0.70 0.05  0.47  0.87  0.40
## FAr            25 18  0.41  0.11  0.40   0.40 0.14  0.23  0.58  0.35
## ADr            26 18  0.77  0.24  0.81   0.80 0.16  0.00  1.00  1.00
## HDr            27 18  0.79  0.24  0.89   0.83 0.13  0.00  1.00  1.00
## RDr            28 17  0.60  0.14  0.62   0.61 0.13  0.33  0.80  0.47
## CRr            29 18  0.29  0.12  0.30   0.29 0.11  0.00  0.50  0.50

```

## C.3 Responses

### C.3.1 Task Time

Mixed-effects poisson regressions, adjusted for overdispersion, found Interface significant in describing task time  $p = .04$ , but the model was not significantly more likely than a FirstBooklet-only model,  $p > .05$ .

```

METIME.glm3.1 <- glmer(Time.min ~ 1 + obs_effect + FirstBooklet + Interface +
(1|obs_effect) + (1 | Participant)
, data=temp, family=poisson)
summary(METIME.glm3.1, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: poisson ( log )
## Formula:
## Time.min ~ 1 + obs_effect + FirstBooklet + Interface + (1 | obs_effect) +
## (1 | Participant)
## Data: temp
##
##      AIC      BIC   logLik deviance df.resid
##  271.5    281.0  -129.8   259.5     30
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.22532 -0.36288 -0.01375  0.29345  1.34619
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## obs_effect (Intercept) 0.03638  0.1907
## Participant (Intercept) 0.02747  0.1658
## Number of obs: 36, groups:  obs_effect, 36; Participant, 18
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.35614    0.14133  23.747 < 2e-16 ***
## obs_effect     -0.02287    0.01148  -1.993  0.04627 *
## FirstBookletTRUE 0.73164    0.22492   3.253  0.00114 **
## InterfaceC+R    0.18104    0.08911   2.032  0.04218 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##              (Intr) obs_ff FBTRUE
## obs_effect  -0.758
## FrstBklTRUE  0.546 -0.918
## InterfacC+R -0.370 -0.002  0.053

```

```

anova(METIME.glm1.1, METIME.glm2.1, METIME.glm3.1)
## Data: temp
## Models:
## METIME.glm1.1: Time.min ~ 1 + obs_effect + (1 | obs_effect) + (1 | Participant)
## METIME.glm2.1: Time.min ~ 1 + obs_effect + FirstBooklet + (1 | obs_effect) +
## METIME.glm2.1: (1 | Participant)
## METIME.glm3.1: Time.min ~ 1 + obs_effect + FirstBooklet + Interface +
## METIME.glm3.1: (1 | obs_effect) + (1 | Participant)
##              Df      AIC      BIC loglik deviance Chisq Chi Df Pr(>Chisq)
## METIME.glm1.1  4 279.41 285.74 -135.70  271.41
## METIME.glm2.1  5 273.30 281.22 -131.65  263.30 8.1031      1  0.004419 **

```

```
## MEltime.glm3.1 6 271.52 281.03 -129.76 259.52 3.7790 1 0.051900 .
##
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### C.3.2 Chart Location

Participants responded most often on the CUSUM chart in both conditions, but the RE charts were almost as popular in the C+R condition.

**Table 29 – Experiment I Response locations for all participants ( $n=18$ ), all scenarios ( $s=5$ ) in each experimental condition (CUSUM only, or CUSUM+RE). Charts are numbered as they appeared on the response forms, from top to bottom.**

Interface Condition	Response Sheet Chart Number								Chart numbers: 1: Heating Driver variable 2: Precipitation Driver variable 3: Consumption 4: Control 5: CUSUM 6: RE – Baseload 7: RE – Heating 8: RE – Precipitation
	1	2	3	4	5	6	7	8	
CUSUM	3	0	9	27	267				
CUSUM+RE	0	0	0	0	104	78	81	65	
Total	3	0	9	27	371	78	81	65	

## C.4 By-Participant Performance

### C.4.1 Detection

Hit rate did not significantly differ between Interface conditions.

```
anova(MEllogit.Q1HR.1, MEllogit.Q1HR.19, MEllogit.Q1HR.20)
## Data: QLevelFactors
## Models:
## MEllogit.Q1HR.1: H ~ (1 | Participant) + (1 | TrueScenario)
## MEllogit.Q1HR.19: H ~ FirstBooklet + (1 | Participant) + (1 | TrueScenario)
## MEllogit.Q1HR.20: H ~ FirstBooklet + Interface + (1 | Participant) + (1 | TrueScenario)
##
##           Df      AIC      BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## MEllogit.Q1HR.1  3 758.24 771.59 -376.12 752.24
## MEllogit.Q1HR.19  4 757.21 775.02 -374.61 749.21 3.0220 1 0.08214 .
## MEllogit.Q1HR.20  5 759.21 781.47 -374.60 749.21 0.0083 1 0.92747
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

However, marked Confidence in a response was slightly correlated  $r(629) = .11, p = .003$  with the response being a Hit. This correlation was significant, and did not differ between interface conditions:

```

MELogit.QlHR.17 <- glmer(H ~ MarkedConfidence
                        + (1 | Participant ) + (1 | TrueScenario)
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.QlHR.17, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: H ~ MarkedConfidence + (1 | Participant) + (1 | TrueScenario)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##  743.4    761.1  -367.7   735.4     627
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.4230 -0.7731  0.3155  0.6918  2.2143
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## Participant (Intercept) 0.2291  0.4787
## TrueScenario (Intercept) 1.2993  1.1399
## Number of obs: 631, groups: Participant, 18; TrueScenario, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.7748    0.5334  -1.453 0.146316
## MarkedConfidence  0.1775    0.0508   3.494 0.000476 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##              (Intr)
## MarkdCnfdnc -0.679

```

```

anova(MELogit.QlHR.1, MELogit.QlHR.17, MELogit.QlHR.18)
## Data: QLevelFactors
## Models:
## MELogit.QlHR.1: H ~ (1 | Participant) + (1 | TrueScenario)
## MELogit.QlHR.17: H ~ MarkedConfidence + (1 | Participant) + (1 | TrueScenario)
## MELogit.QlHR.18: H ~ MarkedConfidence * Interface + (1 | Participant) + (1 |
TrueScenario)
##
##      Df      AIC      BIC loglik deviance  Chisq Chi Df Pr(>Chisq)
## MELogit.QlHR.1  3 758.24 771.59 -376.12  752.24
## MELogit.QlHR.17  4 743.36 761.14 -367.68  735.36 16.8803      1 3.981e-05 ***
## MELogit.QlHR.18  6 747.11 773.80 -367.56  735.11  0.2411      2  0.8864
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

## C.4.2 Diagnosis

Mixed effects logistic regression models reverse fit by and compared by maximum likelihood found right diagnosis rate not significantly associated with experimental order, participant school, first booklet, or Interface condition.

```
anova(MELogit.Q1RDr1, MELogit.Q1RDr16, MELogit.Q1RDr17, MELogit.Q1RDr18,
MELogit.Q1RDr19)
## Data: QLevelFactors
## Models:
## MELogit.Q1RDr1: RightDiag_likely ~ (1 | TrueScenario)
## MELogit.Q1RDr16: RightDiag_likely ~ Order + (1 | TrueScenario)
## MELogit.Q1RDr17: RightDiag_likely ~ Interface + Order + (1 | TrueScenario)
## MELogit.Q1RDr18: RightDiag_likely ~ Interface + FirstBooklet + Order + (1 |
TrueScenario)
## MELogit.Q1RDr19: RightDiag_likely ~ Interface + FirstBooklet + School + Order +
## MELogit.Q1RDr19: (1 | TrueScenario)
##          Df    AIC    BIC loglik deviance  Chisq Chi Df Pr(>Chisq)
## MELogit.Q1RDr1  2 372.35 379.54 -184.17  368.35
## MELogit.Q1RDr16  5 371.34 389.31 -180.67  361.34 7.0109      3  0.07155 .
## MELogit.Q1RDr17  6 371.81 393.38 -179.91  359.81 1.5228      1  0.21720
## MELogit.Q1RDr18  7 372.42 397.59 -179.21  358.42 1.3894      1  0.23851
## MELogit.Q1RDr19  8 374.28 403.04 -179.14  358.28 0.1430      1  0.70532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## C.5 By-Change Type Performance

### C.5.1 Detection

Many factors were considered in reverse fitting a mixed effects logistic regression model to explain likelihood of a change being detected. However, models with interaction effects may not be generalizable, as each combination of factors would be based on just one or two changes.

Two candidate models with main effects only (#13 and 14) were considered.

```
print(anova(MELogit.Q1Det1,MELogit.Q1Det10,MELogit.Q1Det13,MELogit.Q1Det14,MELogit.Q1
Det15,MELogit.Q1Det16,MELogit.Q1Det17,MELogit.Q1Det18,MELogit.Q1Det19,MELogit.Q1Det20
))
## Data: QLevelLogit
## Models:
## MELogit.Q1Det1: Hit ~ (1 | Participant) + (1 | TrueScenario)
## MELogit.Q1Det10: Hit ~ SizeLarge + Evidence + Leading + (1 | Participant) + (1 |
## MELogit.Q1Det10: TrueScenario)
## MELogit.Q1Det13: Hit ~ Cause + SizeLarge + Leading + (1 | Participant) + (1 |
## MELogit.Q1Det13: TrueScenario)
## MELogit.Q1Det14: Hit ~ Cause + SizeLarge + Evidence + Leading + (1 | Participant)
+
## MELogit.Q1Det14: (1 | TrueScenario)
## MELogit.Q1Det15: Hit ~ Interface + Cause + SizeLarge + Evidence + Leading + (1 |
```

```

## MLogit.QlDet15: Participant) + (1 | TrueScenario)
## MLogit.QlDet16: Hit ~ Interface + Cause + SizeLarge + Evidence + Leading +
Counteracting +
## MLogit.QlDet16: (1 | Participant) + (1 | TrueScenario)
## MLogit.QlDet17: Hit ~ Cause + SizeLarge + Evidence + Leading + Interface *
Counteracting +
## MLogit.QlDet17: (1 | Participant) + (1 | TrueScenario)
## MLogit.QlDet18: Hit ~ FirstBooklet + Cause + SizeLarge + Evidence + Leading +
## MLogit.QlDet18: Interface * Counteracting + (1 | Participant) + (1 |
TrueScenario)
## MLogit.QlDet19: Hit ~ Interface * FirstBooklet + Cause + SizeLarge + Evidence +
## MLogit.QlDet19: Leading + Interface * Counteracting + (1 | Participant) +
## MLogit.QlDet19: (1 | TrueScenario)
## MLogit.QlDet20: Hit ~ Interface * FirstBooklet + Interface * Cause + SizeLarge +
## MLogit.QlDet20: Evidence + Leading + Interface * Counteracting + (1 |
Participant) +
## MLogit.QlDet20: (1 | TrueScenario)
##          Df      AIC      BIC logLik deviance  Chisq Chi Df      Pr(>Chisq)
## MLogit.QlDet1  3 661.31 674.18 -327.65  655.31
## MLogit.QlDet10 6 619.22 644.97 -303.61  607.22 48.0849      3 2.043e-10 ***
## MLogit.QlDet13 7 597.12 627.16 -291.56  583.12 24.1063      1 9.116e-07 ***
## MLogit.QlDet14 8 593.36 627.70 -288.68  577.36  5.7521      1 0.016469 *
## MLogit.QlDet15 9 594.47 633.09 -288.24  576.47  0.8934      1 0.344551
## MLogit.QlDet16 10 596.16 639.08 -288.08  576.16  0.3099      1 0.577769
## MLogit.QlDet17 11 588.21 635.42 -283.11  566.21  9.9504      1 0.001608 **
## MLogit.QlDet18 12 589.45 640.95 -282.73  565.45  0.7600      1 0.383343
## MLogit.QlDet19 13 586.94 642.73 -280.47  560.94  4.5099      1 0.033699 *
## MLogit.QlDet20 15 588.92 653.30 -279.46  558.92  2.0160      2 0.364956
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

MLogit.QlDet14 <- glmer(Hit ~ Cause
                        + SizeLarge
                        + Evidence
                        + Leading
                        + (1 | Participant) + (1 | TrueScenario)
                        , data = QLevelLogit, family = binomial, nAGQ = 1)
summary(MLogit.QlDet14, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula:
## Hit ~ Cause + SizeLarge + Evidence + Leading + (1 | Participant) +
## (1 | TrueScenario)
## Data: QLevelLogit
##
##          AIC      BIC  logLik deviance df.resid
##    593.4    627.7  -288.7   577.4     532
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.1483 -0.7670  0.3162  0.6417  1.7012

```

```

##
## Random effects:
## Groups      Name          Variance Std.Dev.
## Participant (Intercept) 0.1228  0.3504
## TrueScenario (Intercept) 1.0486  1.0240
## Number of obs: 540, groups: Participant, 18; TrueScenario, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.73158   0.68611  -1.066  0.28630
## CauseB       1.58274   0.32677   4.844 1.27e-06 ***
## CauseC       0.67718   0.47693   1.420  0.15565
## SizeLargeTRUE 1.38912   0.27140   5.118 3.08e-07 ***
## Evidence     0.08954   0.03810   2.350  0.01876 *
## LeadingTRUE  -0.70611   0.25217  -2.800  0.00511 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##              (Intr) CauseB CauseC SLTRUE Evidnc
## CauseB      -0.535
## CauseC      -0.769  0.582
## SizeLrgTRUE  0.118  0.119 -0.259
## Evidence    -0.779  0.451  0.852 -0.249
## LeadingTRUE  0.028 -0.100 -0.170 -0.204 -0.242

```

```

MELogit.Q1Det13 <- glmer(Hit ~ Cause
                        + SizeLarge
                        + Leading
                        + (1 | Participant) + (1 | TrueScenario)
                        , data = QLevelLogit, family = binomial, nAGQ = 1)
summary(MELogit.Q1Det13, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: Hit ~ Cause + SizeLarge + Leading + (1 | Participant) + (1 |
## TrueScenario)
## Data: QLevelLogit
##
##      AIC      BIC   logLik deviance df.resid
## 597.1    627.2  -291.6   583.1     533
##
## Scaled residuals:
##      Min      1Q   Median      3Q      Max
## -3.5761 -0.7383  0.3496  0.6246  1.8203
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## Participant (Intercept) 0.1228  0.3504
## TrueScenario (Intercept) 1.2037  1.0971
## Number of obs: 540, groups: Participant, 18; TrueScenario, 10

```

```
##
## Fixed effects:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.5556    0.4495   1.236  0.2164
## CauseB      1.2801    0.2838   4.511 6.46e-06 ***
## CauseC     -0.2907    0.2515  -1.156  0.2477
## SizeLargeTRUE 1.5989    0.2541   6.292 3.13e-10 ***
## LeadingTRUE  -0.5856    0.2455  -2.385  0.0171 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##           (Intr) CauseB CauseC SLTRUE
## CauseB      -0.329
## CauseC     -0.301  0.448
## SizeLrgTRUE -0.122  0.203 -0.081
## LeadingTRUE -0.267  0.029  0.047 -0.272
```

Forward fitting more complex random effect structures (with TrueScenario) did not significantly improve either model.

## C.5.2 Diagnosis

As for diagnosis, several mixed effects logistic regression models were fit to describe probability of right diagnosis of a detected, diagnosed change. The most likely model described diagnosis in terms of an interaction between change cause and Interface type.

```
MELogit.QlDiag9 <- glmer(RightDiag ~ Interface * Cause
                        + (1|Participant)
                        , data = QLevelLogit, family = binomial, nAGQ = 1)
summary(MELogit.QlDiag9, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: RightDiag ~ Interface * Cause + (1 | Participant)
## Data: QLevelLogit
##
##           AIC      BIC   logLik deviance df.resid
##    345.3    370.5  -165.7   331.3     262
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.0827 -0.9912  0.5161  0.8668  2.6245
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## Participant (Intercept) 0.115    0.3391
## Number of obs: 269, groups: Participant, 18
##
```

```

## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.3169    0.4217   3.123 0.001791 **
## InterfaceC+R  -1.2334    0.5074  -2.431 0.015072 *
## CauseB         -1.1576    0.4897  -2.364 0.018080 *
## CauseC         -2.8999    0.6502  -4.460 8.19e-06 ***
## InterfaceC+R:CauseB  2.1799    0.6495   3.356 0.000789 ***
## InterfaceC+R:CauseC  2.7149    0.7812   3.475 0.000511 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##              (Intr) IntC+R CauseB CauseC IC+R:CB
## InterfacC+R -0.800
## CauseB      -0.820  0.681
## CauseC      -0.630  0.524  0.535
## IntrfC+R:CB  0.626 -0.782 -0.759 -0.411
## IntrfC+R:CC  0.512 -0.644 -0.438 -0.824  0.504

```

```

print(anova(MELogit.QlDiag1, MELogit.QlDiag7, MELogit.QlDiag8, MELogit.QlDiag9,
MELogit.QlDiag10, MELogit.QlDiag11, MELogit.QlDiag12, MELogit.QlDiag13,
MELogit.QlDiag14,MELogit.QlDiag15,MELogit.QlDiag16,MELogit.QlDiag17,MELogit.QlDiag18,
MELogit.QlDiag19))
## Data: QLevelLogit
## Models:
## MELogit.QlDiag1: RightDiag ~ (1 | Participant)
## MELogit.QlDiag7: RightDiag ~ Cause + (1 | Participant)
## MELogit.QlDiag8: RightDiag ~ Interface + Cause + (1 | Participant)
## MELogit.QlDiag9: RightDiag ~ Interface * Cause + (1 | Participant)
## MELogit.QlDiag10: RightDiag ~ Interface * Cause + Counteracting + (1 |
Participant)
## MELogit.QlDiag11: RightDiag ~ FirstBooklet + Interface * Cause + Counteracting +
## MELogit.QlDiag11:      (1 | Participant)
## MELogit.QlDiag12: RightDiag ~ FirstBooklet + Interface * Cause + Evidence +
Counteracting +
## MELogit.QlDiag12:      (1 | Participant)
## MELogit.QlDiag13: RightDiag ~ FirstBooklet + Interface * Cause + SizeLarge +
Evidence +
## MELogit.QlDiag13:      Counteracting + (1 | Participant)
## MELogit.QlDiag14: RightDiag ~ FirstBooklet + Leading + Interface * Cause +
SizeLarge +
## MELogit.QlDiag14:      Evidence + Counteracting + (1 | Participant)
## MELogit.QlDiag15: RightDiag ~ FirstBooklet + Interface * Leading + Interface *
## MELogit.QlDiag15:      Cause + SizeLarge + Evidence + Counteracting + (1 |
Participant)
## MELogit.QlDiag16: RightDiag ~ Interface * FirstBooklet + Interface * Leading +
## MELogit.QlDiag16:      Interface * Cause + SizeLarge + Evidence + Counteracting +
## MELogit.QlDiag16:      (1 | Participant)
## MELogit.QlDiag17: RightDiag ~ Interface * FirstBooklet + Interface * Leading +
## MELogit.QlDiag17:      Interface * Cause + SizeLarge + Interface * Evidence +
Counteracting +

```

```

## MLogit.QlDiag17: (1 | Participant)
## MLogit.QlDiag18: RightDiag ~ Interface * FirstBooklet + Interface * Leading +
## MLogit.QlDiag18: Interface * Cause + Interface * SizeLarge + Interface *
Evidence +
## MLogit.QlDiag18: Counteracting + (1 | Participant)
## MLogit.QlDiag19: RightDiag ~ Interface * FirstBooklet + Interface * Leading +
## MLogit.QlDiag19: Interface * Cause + Interface * SizeLarge + Interface *
Evidence +
## MLogit.QlDiag19: Interface * Counteracting + (1 | Participant)
##      Df      AIC      BIC  loglik deviance  Chisq Chi Df      Pr(>Chisq)
## MLogit.QlDiag1  2 371.99 379.18 -183.99  367.99
## MLogit.QlDiag7  4 358.41 372.79 -175.20  350.41 17.5779      2  0.0001524 ***
## MLogit.QlDiag8  5 358.32 376.29 -174.16  348.32  2.0933      1  0.1479490
## MLogit.QlDiag9  7 345.32 370.48 -165.66  331.32 16.9992      2  0.0002035 ***
## MLogit.QlDiag10 8 344.12 372.88 -164.06  328.12  3.1932      1  0.0739453 .
## MLogit.QlDiag11 9 345.29 377.64 -163.65  327.29  0.8306      1  0.3620874
## MLogit.QlDiag12 10 346.91 382.85 -163.45  326.91  0.3861      1  0.5343396
## MLogit.QlDiag13 11 348.38 387.92 -163.19  326.38  0.5287      1  0.4671428
## MLogit.QlDiag14 12 350.14 393.28 -163.07  326.14  0.2369      1  0.6264264
## MLogit.QlDiag15 13 349.87 396.60 -161.94  323.87  2.2678      1  0.1320881
## MLogit.QlDiag16 14 350.69 401.02 -161.35  322.69  1.1801      1  0.2773378
## MLogit.QlDiag17 15 352.47 406.39 -161.24  322.47  0.2223      1  0.6373192
## MLogit.QlDiag18 16 354.39 411.91 -161.20  322.39  0.0781      1  0.7798392
## MLogit.QlDiag19 17 356.37 417.48 -161.19  322.37  0.0204      1  0.8863958
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Contrasts showed that this model found inconsistent effects of which change types were harder to diagnose.

```

MLogit.QlDiag9.c1 <- rbind("Interface C, Change A vs B" = c(0,0,1,0,0,0),
                          "Interface C, Change A vs C" = c(0,0,0,1,0,0),
                          "Interface C, Change B vs C" = c(0,0,-1,1,0,0),

                          "Interface C+R, Change A vs B" = c(0,0,1,0,1,0),
                          "Interface C+R, Change A vs C" = c(0,0,0,1,0,1),
                          "Interface C+R, Change B vs C" = c(0,0,-1,1,-1,1)
)

summary(glht(MLogit.QlDiag9, MLogit.QlDiag9.c1))
##
## Simultaneous Tests for General Linear Hypotheses
##
## Fit: glmer(formula = RightDiag ~ Interface * Cause + (1 | Participant),
## data = QLevelLogit, family = binomial, nAGQ = 1)
##
## Linear Hypotheses:
##
##      Estimate Std. Error z value Pr(>|z|)
## Interface C, Change A vs B == 0 -1.1576  0.4897 -2.364  0.0917 .
## Interface C, Change A vs C == 0 -2.8999  0.6502 -4.460 <1e-04 ***
## Interface C, Change B vs C == 0 -1.7423  0.5672 -3.072  0.0119 *

```

```
## Interface C+R, Change A vs B == 0  1.0223    0.4230   2.417   0.0802 .
## Interface C+R, Change A vs C == 0  -0.1850    0.4427  -0.418   0.9915
## Interface C+R, Change B vs C == 0  -1.2073    0.4537  -2.661   0.0416 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

A second contrast showed mixed performance effects of the C+R interface condition:

```
MELogit.QlDiag9.c2 <- rbind("Change A, Interface C vs C+R" = c(0,1,0,0,0,0),
                          "Change B, Interface C vs C+R" = c(0,1,0,0,1,0), #
                          "Change C, Interface C vs C+R" = c(0,1,0,0,0,1)
)

summary(glht(MELogit.QlDiag9, MELogit.QlDiag9.c2))
##
## Simultaneous Tests for General Linear Hypotheses
##
## Fit: glmer(formula = RightDiag ~ Interface * Cause + (1 | Participant),
##   data = QLevelLogit, family = binomial, nAGQ = 1)
##
## Linear Hypotheses:
##
##           Estimate Std. Error z value Pr(>|z|)
## Change A, Interface C vs C+R == 0  -1.2334    0.5074  -2.431   0.0445 *
## Change B, Interface C vs C+R == 0   0.9465    0.4046   2.339   0.0568 .
## Change C, Interface C vs C+R == 0   1.4815    0.5979   2.478   0.0391 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```



## Appendix D Experiment II Statistical Output

Data analysis was performed with the R statistical language, version 3.0.2 (R Project Team, n.d.).

### D.1 Experimental Design

#### D.1.1 Model training data

The weather data used for the training set was drawn from Environment Canada Toronto Pearson Airport records (Weather Underground, n.d.). The cut-in break point for the heating system was arbitrarily set at below 14C (with 10% introduced error). Holidays were counted as weekends. The “Generator” driver was entirely synthetic (see Section 5.2.1). Below are summary statistics of the first year’s data (starting July 2009) used to train the baseline model.

```
> describe(OutputFull[which(OutputFull$Timestamp <
as.Date("2010/07/01")),)][,c("vars", "n", "mean", "sd", "median", "min", "max", "range", "se"
)]
```

	vars	n	mean	sd	median	min	max	range	se
Timestamp*	1	365	14608.00	105.51	14608.00	14426.00	14790.00	364.00	5.52
Temperature	2	365	9.19	9.68	10.00	-14.00	27.00	41.00	0.51
IsWorkday	4	365	0.69	0.46	1.00	0.00	1.00	1.00	0.02
Generator	5	365	5.87	6.81	0.00	0.00	23.40	23.40	0.36
HDD	6	365	5.60	6.86	2.00	0.00	26.00	26.00	0.36
HDDTrue	8	365	5.65	6.85	2.25	0.00	26.45	26.45	0.36
Consumption	9	365	14464.63	6403.96	14257.11	3011.53	35535.98	32524.44	335.20
Base	10	365	3468.45	357.62	3459.74	2501.87	4578.93	2077.06	18.72
HDDCoef	11	365	623.41	64.57	621.95	467.52	881.34	413.82	3.38
GenCoef	12	365	618.46	60.04	620.94	469.05	804.68	335.63	3.14
HDDBreak	13	365	12.05	1.21	12.10	8.33	15.86	7.54	0.06
WorkdayCoef	15	365	5098.24	510.37	5099.47	3435.57	6379.18	2943.62	26.71
WeibullError	16	365	293.19	198.84	243.68	11.89	1162.56	1150.67	10.41

### D.2 Data Summaries from Experiment II

#### D.2.1 Performance Data, Aggregated by-Participant

Below is a summary of the Experiment II metrics aggregated at Participant-level. Measures ending with \_inbox and \_likely refer to the scoring rules discussed in Section 5.2.8.

```
> print(describe(PLevelFactors))
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
FirstBooklet*	1	66	0.50	0.50	0.50	0.50	0.74	0.00	1.00	1.00	0.00
Participant*	2	66	17.00	9.59	17.00	17.00	11.86	1.00	33.00	32.00	0.00
Order*	3	66	2.45	1.14	2.00	2.44	1.48	1.00	4.00	3.00	0.05

School*	4	66	1.33	0.48	1.00	1.30	0.00	1.00	2.00	1.00	0.69
Time.min	5	66	42.52	15.61	38.50	40.96	15.57	19.00	91.00	72.00	0.85
TrueScenarios*	6	66	1.50	0.50	1.50	1.50	0.74	1.00	2.00	1.00	0.00
Interface*	7	66	1.50	0.50	1.50	1.50	0.74	1.00	2.00	1.00	0.00
Responses	8	66	20.14	6.48	19.50	19.63	6.67	9.00	39.00	30.00	0.74
TrueChanges	9	66	13.00	0.00	13.00	13.00	0.00	13.00	13.00	0.00	NaN
MarkedCause	10	66	16.23	6.81	16.00	16.06	6.67	1.00	32.00	31.00	0.24
Change_inbox	11	66	6.21	2.41	6.00	6.11	2.97	2.00	11.00	9.00	0.30
Diag_inbox	12	66	5.09	2.45	5.00	5.00	2.97	0.00	11.00	11.00	0.35
RightDiag_inbox	13	66	1.65	1.36	1.00	1.54	1.48	0.00	5.00	5.00	0.60
Change_likely	14	66	8.73	2.04	8.00	8.72	2.97	5.00	13.00	8.00	0.06
Diag_likely	15	66	7.08	2.59	7.00	7.13	2.97	0.00	12.00	12.00	-0.31
RightDiag_likely	16	66	2.47	1.49	2.00	2.46	1.48	0.00	7.00	7.00	0.30
H	17	66	8.73	2.04	8.00	8.72	2.97	5.00	13.00	8.00	0.06
FA	18	66	11.41	5.21	11.00	10.93	4.45	4.00	28.00	24.00	0.85
M	19	66	4.27	2.04	5.00	4.28	2.97	0.00	8.00	8.00	-0.06
AD	20	66	16.23	6.81	16.00	16.06	6.67	1.00	32.00	31.00	0.24
HD	21	66	7.08	2.59	7.00	7.13	2.97	0.00	12.00	12.00	-0.31
RD	22	66	2.47	1.49	2.00	2.46	1.48	0.00	7.00	7.00	0.30
WD	23	66	4.61	2.05	5.00	4.59	1.48	0.00	9.00	9.00	0.02
Hr	24	66	0.67	0.16	0.62	0.67	0.23	0.38	1.00	0.62	0.06
FAr	25	66	0.55	0.10	0.56	0.55	0.11	0.29	0.72	0.44	-0.51
ADr	26	66	0.80	0.19	0.83	0.83	0.18	0.05	1.00	0.95	-1.50
HDr	27	66	0.80	0.21	0.85	0.84	0.23	0.00	1.00	1.00	-1.37
RDr	28	65	0.35	0.18	0.33	0.34	0.20	0.00	0.80	0.80	0.22
CRr	29	66	0.13	0.07	0.12	0.12	0.06	0.00	0.29	0.29	0.22

## D.2.2 Performance Data, by Participant and Interface

```
print(describeBy(PLevelFactors, PLevelFactors$Interface))

## group: C
##          vars  n  mean   sd median trimmed  mad  min  max
## FirstBooklet*    1 33  0.52 0.51  1.00   0.52 0.00 0.00 1.00
## Participant*    2 33 17.00 9.67 17.00 17.00 11.86 1.00 33.00
## Order*          3 33  2.45 1.15  2.00   2.44 1.48 1.00 4.00
## School*         4 33  1.33 0.48  1.00   1.30 0.00 1.00 2.00
## Time.min        5 33 40.03 14.96 36.00 38.56 14.83 20.00 80.00
## TrueScenarios*  6 33  1.48 0.51  1.00   1.48 0.00 1.00 2.00
## Interface*      7 33  1.00 0.00  1.00   1.00 0.00 1.00 1.00
## Responses       8 33 20.21 6.55 19.00 19.74 5.93 11.00 37.00
## TrueChanges     9 33 13.00 0.00 13.00 13.00 0.00 13.00 13.00
## MarkedCause    10 33 15.42 7.26 16.00 15.33 8.90 1.00 28.00
## Change_inbox   11 33  6.24 2.36  6.00   6.19 2.97 2.00 11.00
## Diag_inbox     12 33  4.67 2.41  4.00   4.59 2.97 0.00 10.00
## RightDiag_inbox 13 33  1.21 1.24  1.00   1.04 1.48 0.00 5.00
## Change_likely  14 33  8.79 2.13  8.00   8.81 2.97 5.00 13.00
## Diag_likely    15 33  6.55 2.84  7.00   6.67 2.97 0.00 11.00
## RightDiag_likely 16 33  1.85 1.30  2.00   1.78 1.48 0.00 5.00
## H              17 33  8.79 2.13  8.00   8.81 2.97 5.00 13.00
## FA             18 33 11.42 5.03 10.00 11.04 4.45 4.00 24.00
## M              19 33  4.21 2.13  5.00   4.19 2.97 0.00 8.00
## AD             20 33 15.42 7.26 16.00 15.33 8.90 1.00 28.00
## HD             21 33  6.55 2.84  7.00   6.67 2.97 0.00 11.00
## RD             22 33  1.85 1.30  2.00   1.78 1.48 0.00 5.00
```

```

## WD                23 33  4.70  2.17  5.00   4.70  2.97  0.00  9.00
## Hr                24 33  0.68  0.16  0.62   0.68  0.23  0.38  1.00
## FAr              25 33  0.55  0.09  0.56   0.56  0.10  0.33  0.67
## ADr              26 33  0.76  0.23  0.77   0.79  0.19  0.05  1.00
## HDr              27 33  0.73  0.24  0.75   0.76  0.21  0.00  1.00
## RDr              28 32  0.28  0.16  0.27   0.28  0.16  0.00  0.57
## CRr              29 33  0.09  0.06  0.09   0.09  0.05  0.00  0.21
## -----
## group: C+R
##                vars  n  mean   sd median trimmed  mad  min  max
## FirstBooklet*   1 33  0.48  0.51  0.00   0.48  0.00  0.00  1.00
## Participant*    2 33 17.00  9.67 17.00  17.00 11.86  1.00 33.00
## Order*          3 33  2.45  1.15  2.00   2.44  1.48  1.00  4.00
## School*         4 33  1.33  0.48  1.00   1.30  0.00  1.00  2.00
## Time.min        5 33 45.00 16.08 39.00  43.56 14.83 19.00 91.00
## TrueScenarios*  6 33  1.52  0.51  2.00   1.52  0.00  1.00  2.00
## Interface*      7 33  2.00  0.00  2.00   2.00  0.00  2.00  2.00
## Responses       8 33 20.06  6.50 20.00  19.52  7.41  9.00 39.00
## TrueChanges     9 33 13.00  0.00 13.00  13.00  0.00 13.00 13.00
## MarkedCause    10 33 17.03  6.34 16.00  16.70  5.93  7.00 32.00
## Change_inbox   11 33  6.18  2.51  5.00   6.04  2.97  3.00 11.00
## Diag_inbox     12 33  5.52  2.46  5.00   5.33  2.97  2.00 11.00
## RightDiag_inbox 13 33  2.09  1.35  2.00   2.07  1.48  0.00  5.00
## Change_likely  14 33  8.67  1.98  9.00   8.63  2.97  5.00 12.00
## Diag_likely    15 33  7.61  2.22  7.00   7.56  2.97  4.00 12.00
## RightDiag_likely 16 33  3.09  1.42  3.00   3.07  1.48  0.00  7.00
## H              17 33  8.67  1.98  9.00   8.63  2.97  5.00 12.00
## FA             18 33 11.39  5.46 11.00  10.81  5.93  4.00 28.00
## M              19 33  4.33  1.98  4.00   4.37  2.97  1.00  8.00
## AD             20 33 17.03  6.34 16.00  16.70  5.93  7.00 32.00
## HD             21 33  7.61  2.22  7.00   7.56  2.97  4.00 12.00
## RD             22 33  3.09  1.42  3.00   3.07  1.48  0.00  7.00
## WD             23 33  4.52  1.95  4.00   4.48  1.48  1.00  8.00
## Hr            24 33  0.67  0.15  0.69   0.66  0.23  0.38  0.92
## FAr           25 33  0.55  0.11  0.55   0.55  0.12  0.29  0.72
## ADr           26 33  0.84  0.13  0.85   0.85  0.16  0.50  1.00
## HDr           27 33  0.88  0.15  0.91   0.89  0.13  0.44  1.00
## RDr           28 33  0.41  0.19  0.43   0.41  0.21  0.00  0.80
## CRr           29 33  0.16  0.07  0.15   0.16  0.08  0.00  0.29

```

### D.2.3 Questionnaire Data, by Interface

```

describeBy(PLevelQuestionnaire, group = PLevelQuestionnaire$Interface)
## group: C
##                vars  n  mean   sd median trimmed  mad  min  max  range  skew
## Participant*    1 33 17.00  9.67   17  17.00 11.86  1 33   32  0.00
## School*         2 33  1.33  0.48    1   1.30  0.00  1  2    1  0.68
## Order*          3 33  2.45  1.15    2   2.44  1.48  1  4    3  0.05
## Easy            4 33  3.58  1.00    4   3.59  1.48  2  5    3 -0.38
## When           5 33  3.70  0.98    4   3.78  0.00  1  5    4 -0.92
## What           6 33  2.82  1.04    3   2.85  1.48  1  5    4 -0.13
## Inform         7 33  3.85  0.80    4   3.89  0.00  2  5    3 -0.46
## Confuse        8 33  2.73  0.94    3   2.74  1.48  1  4    3  0.11
## Interface*     9 33  1.00  0.00    1   1.00  0.00  1  1    0  NaN

```

```
## -----
## group: C+R
##          vars  n  mean   sd median trimmed   mad min max range  skew
## Participant*  1 33 17.00 9.67   17  17.00 11.86   1 33  32  0.00
## School*       2 33  1.33 0.48    1   1.30  0.00   1  2   1  0.68
## Order*        3 33  2.45 1.15    2   2.44  1.48   1  4   3  0.05
## Easy          4 33  3.67 0.89    4   3.74  0.00   1  5   4 -1.14
## When          5 33  3.82 0.77    4   3.89  0.00   2  5   3 -0.90
## What          6 32  3.66 0.94    4   3.73  0.00   1  5   4 -0.90
## Inform        7 33  3.94 0.61    4   3.93  0.00   3  5   2  0.02
## Confuse       8 33  2.27 0.76    2   2.26  0.00   1  4   3  0.35
## Interface*    9 33  2.00 0.00    2   2.00  0.00   2  2   0  NaN
```

## D.3 Response modes

### D.3.1 Task Time

Below are the statistical outputs underlying Section 5.5.2.3.

```
summary(MEtime.glmer3.1, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: poisson ( log )
## Formula:
## Time.min ~ 1 + obs_effect + FirstBooklet + Interface + (1 | obs_effect) +
## (1 | Participant)
## Data: temp
##
##          AIC      BIC   logLik deviance df.resid
##    514.0    527.1  -251.0    502.0     60
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.23446 -0.40045 -0.04766  0.36597  1.15202
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## obs_effect (Intercept) 0.04863  0.2205
## Participant (Intercept) 0.00000  0.0000
## Number of obs: 66, groups:  obs_effect, 66; Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.3918213  0.0853327  39.75 < 2e-16 ***
## obs_effect    0.0009589  0.0034943   0.27  0.78377
## FirstBookletTRUE 0.4172093  0.1327885   3.14  0.00168 **
## InterfaceC+R    0.1314125  0.0668073   1.97  0.04918 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##              (Intr) obs_ff FBTRUE
```

```
## obs_effect -0.706
## FrstBklTRUE 0.391 -0.863
## InterfacC+R -0.417 0.006 0.010
```

```
anova(MEtime.glmer1.1, MEtime.glmer2.1, MEtime.glmer3.1)
## Data: temp
## Models:
## MEtime.glmer1.1: Time.min ~ 1 + obs_effect + (1 | obs_effect) + (1 | Participant)
## MEtime.glmer2.1: Time.min ~ 1 + obs_effect + FirstBooklet + (1 | obs_effect) +
## MEtime.glmer2.1:      (1 | Participant)
## MEtime.glmer3.1: Time.min ~ 1 + obs_effect + FirstBooklet + Interface + (1 |
obs_effect) +
## MEtime.glmer3.1:      (1 | Participant)
##          Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## MEtime.glmer1.1 4 522.22 530.97 -257.11  514.22
## MEtime.glmer2.1 5 515.72 526.67 -252.86  505.72 8.4918      1 0.003567 **
## MEtime.glmer3.1 6 513.96 527.10 -250.98  501.96 3.7671      1 0.052271 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The MEtime.glmer3.1 model has all significant variables, but better explains the trial time only at  $p = .052$ .

## D.4 By-Participant M&T Performance

### D.4.1 Interface and Order Effects on Detection

Effects of experimental conditions and orders were not considered practically significant (see Section 5.5.3). However, the effects were verified with reverse-fit mixed-effects logit models.

```
MELogit.Q1HR.1 <- glmer(H ~ (1 | Participant ) + (1 | TrueScenario)
, data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1HR.1, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: H ~ (1 | Participant) + (1 | TrueScenario)
## Data: QLevelFactors
##
##          AIC      BIC   logLik deviance df.resid
## 1724.4 1739.9 -859.2 1718.4 1326
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.6409 -0.8226 -0.5464  0.9273  1.9843
##
## Random effects:
```

```
## Groups      Name      Variance Std.Dev.
## Participant (Intercept) 0.02243  0.1498
## TrueScenario (Intercept) 0.47773  0.6912
## Number of obs: 1329, groups: Participant, 33; TrueScenario, 10
##
## Fixed effects:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.1890      0.2288  -0.826   0.409
```

```
MELogit.Q1HR.20 <- glmer(H ~ FirstBooklet
                        + Interface
                        + (1 | Participant ) + (1 | TrueScenario)
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1HR.20, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula:
## H ~ FirstBooklet + Interface + (1 | Participant) + (1 | TrueScenario)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
## 1727.0  1753.0  -858.5  1717.0    1324
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.6614 -0.8239 -0.5498  0.9313  2.0395
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## Participant (Intercept) 0.02201  0.1484
## TrueScenario (Intercept) 0.48079  0.6934
## Number of obs: 1329, groups: Participant, 33; TrueScenario, 10
##
## Fixed effects:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.117075   0.243944  -0.480   0.631
## FirstBookletTRUE -0.134739   0.116797  -1.154   0.249
## InterfaceC+R    -0.007471   0.117043  -0.064   0.949
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##           (Intr) FBTRUE
## FrstBklTRUE -0.242
## InterfacC+R -0.238  0.000
```

```
MELogit.Q1HR.19 <- glmer(H ~ FirstBooklet
                        + (1 | Participant ) + (1 | TrueScenario)
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1HR.19, corr = FALSE)
```

```

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: H ~ FirstBooklet + (1 | Participant) + (1 | TrueScenario)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
## 1725.0  1745.8  -858.5  1717.0    1325
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.6582 -0.8228 -0.5494  0.9293  2.0357
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## Participant (Intercept) 0.0220  0.1483
## TrueScenario (Intercept) 0.4808  0.6934
## Number of obs: 1329, groups: Participant, 33; TrueScenario, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.1208    0.2369  -0.510   0.610
## FirstBookletTRUE -0.1347    0.1168  -1.154   0.249
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##              (Intr)
## FrstBklTRUE -0.249

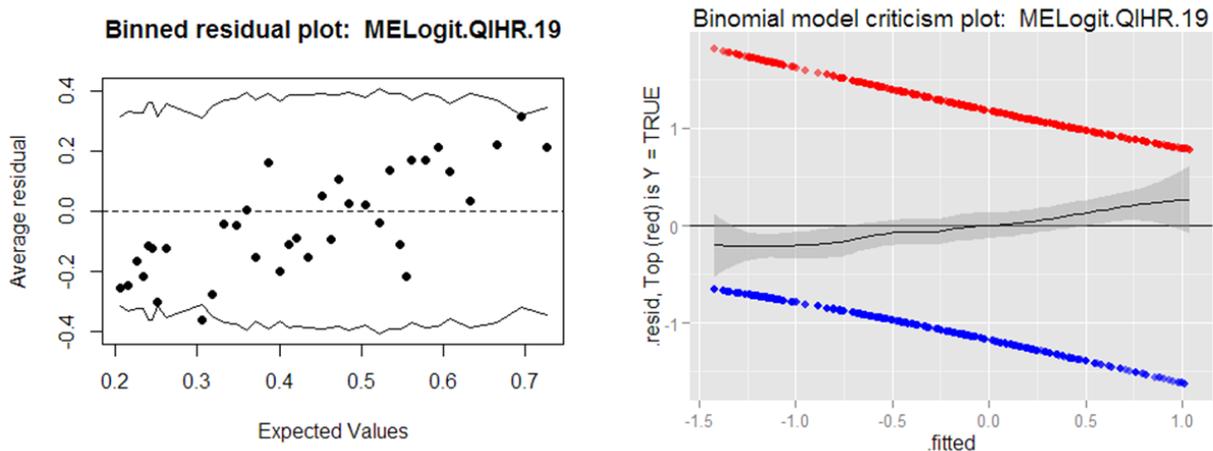
```

```

anova(MELogit.Q1HR.1, MELogit.Q1HR.19, MELogit.Q1HR.20)
## Data: QLevelFactors
## Models:
## MELogit.Q1HR.1: H ~ (1 | Participant) + (1 | TrueScenario)
## MELogit.Q1HR.19: H ~ FirstBooklet + (1 | Participant) + (1 | TrueScenario)
## MELogit.Q1HR.20: H ~ FirstBooklet + Interface + (1 | Participant) + (1 |
TrueScenario)
##              Df      AIC      BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## MELogit.Q1HR.1  3 1724.4 1739.9 -859.18  1718.4
## MELogit.Q1HR.19  4 1725.0 1745.8 -858.51  1717.0 1.3289      1    0.2490
## MELogit.Q1HR.20  5 1727.0 1753.0 -858.51  1717.0 0.0041      1    0.949

```

No evidence was found that FirstBooklet or Interface have a causal effect on likelihood of any given response being a hit.



**Figure 51 – Residual inspection plots for model MELogit.QIHR.19, explaining likelihood of a response being a hit. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.**

#### D.4.2 Interface Effect on Attempted Diagnoses Mixed Effect Logit Models

This appendix summarizes the models used to test whether participants were more likely to mark causes for any given response (Attempted Diagnosis AD<sub>r</sub> in Table 17), described in Section 5.5.3.2. First, simple and more complex models were compared by likelihood ratio tests:

```
## Correcting for number of responses, were participants more likely to make a guess
## in the C+R condition?
##           Use binomial model? But each line is a response, so how to calculate
## offset?
##           Do not need to calculate offset, since each line has an existence of 1

MELogit.QIADr1 <- glmer(AD ~ 1 + (1|Participant) , data=QLevelFactors,
family=binomial)
print(summary(MELogit.QIADr1))
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: AD ~ 1 + (1 | Participant)
## Data: QLevelFactors
##
##           AIC           BIC    logLik deviance df.resid
##    1221.1    1231.5   -608.6   1217.1    1327
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.5750  0.2186  0.3540  0.5032  1.0775
##
## Random effects:
##  Groups      Name          Variance Std.Dev.
## Participant (Intercept) 0.9168   0.9575
## Number of obs: 1329, groups: Participant, 33
```

```
##
## Fixed effects:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.6589    0.1888   8.785  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MELogit.QLADr19 <- glmer(AD ~ 1 + FirstBooklet * Interface + (1|Participant),
data=QLevelFactors, family=binomial)
print(summary(MELogit.QLADr19))
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: AD ~ 1 + FirstBooklet * Interface + (1 | Participant)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
## 1202.7  1228.7  -596.4  1192.7    1324
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -6.1879  0.2066  0.3411  0.5057  1.3727
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## Participant (Intercept) 0.9397  0.9694
## Number of obs: 1329, groups: Participant, 33
##
## Fixed effects:
##
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.2243    0.2834  4.320 1.56e-05 ***
## FirstBookletTRUE    0.3341    0.3970  0.841  0.400
## InterfaceC+R       0.4948    0.4004  1.236  0.217
## FirstBookletTRUE:InterfaceC+R 0.2574    0.7566  0.340  0.734
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) FrBTRUE IntC+R
## FrstBklTRUE -0.706
## InterfacC+R -0.700  0.870
## FBTRUE:IC+R  0.657 -0.923 -0.925
```

```
MELogit.QLADr18 <- glmer(AD ~ 1 + FirstBooklet + Interface + Order + (1|Participant),
data=QLevelFactors, family=binomial)
print(summary(MELogit.QLADr18))
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: AD ~ 1 + FirstBooklet + Interface + Order + (1 | Participant)
```

```

## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
## 1206.2  1242.5   -596.1  1192.2    1322
##
## Scaled residuals:
##      Min      1Q  Median      3Q      Max
## -6.3594  0.2064  0.3457  0.5141  1.3827
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## Participant (Intercept) 0.9229  0.9607
## Number of obs: 1329, groups: Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.1290    0.3786   2.982  0.00286 **
## FirstBookletTRUE  0.4607    0.1529   3.014  0.00258 **
## InterfaceC+R      0.6202    0.1528   4.058 4.95e-05 ***
## OrderB            0.2989    0.5263   0.568  0.57010
## OrderC           -0.0687    0.5222  -0.132  0.89532
## OrderD           -0.1008    0.5221  -0.193  0.84691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) FBTRUE IntC+R OrderB OrderC
## FrstBklTRUE -0.213
## InterfacC+R -0.227  0.111
## OrderB      -0.664  0.031  0.041
## OrderC      -0.657 -0.029  0.027  0.471
## OrderD      -0.680  0.051  0.060  0.475  0.476

```

```

MELogit.QLADr17 <- glmer(AD ~ 1 + FirstBooklet + Interface + (1|Participant),
data=QLevelFactors, family=binomial)
print(summary(MELogit.QLADr17))
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: AD ~ 1 + FirstBooklet + Interface + (1 | Participant)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
## 1200.8  1221.6   -596.4  1192.8    1325
##
## Scaled residuals:
##      Min      1Q  Median      3Q      Max
## -6.1193  0.2034  0.3426  0.5079  1.3759
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## Participant (Intercept) 0.9452  0.9722
## Number of obs: 1329, groups: Participant, 33

```

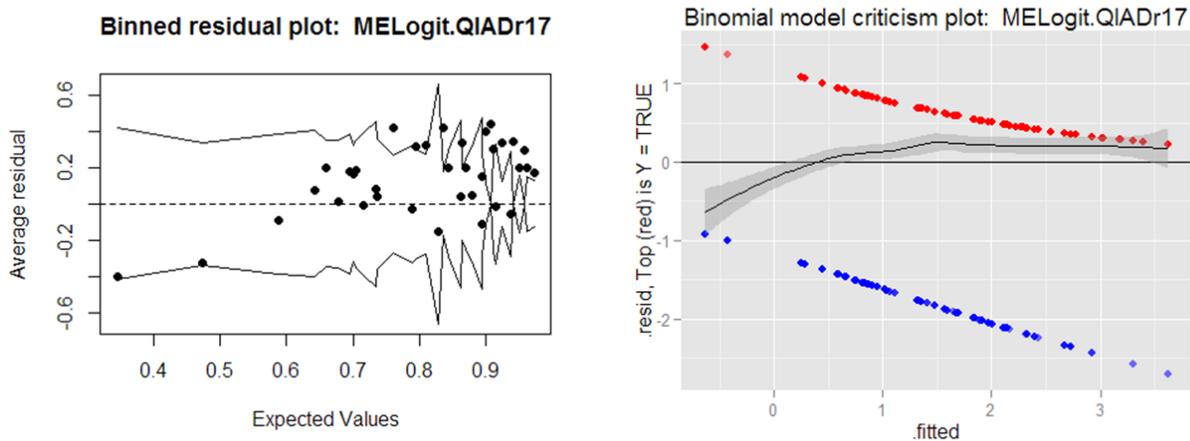
```
##
## Fixed effects:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.1616    0.2140   5.429 5.67e-08 ***
## FirstBookletTRUE 0.4589    0.1522   3.014 0.00258 **
## InterfaceC+R   0.6208    0.1524   4.072 4.65e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) FBTRUE
## FrstBklTRUE -0.342
## InterfacC+R -0.321  0.102
```

Comparing the simplest model (participant random effects only) to the most complex (Learning, Order, and Interface effects):

```
anova(MELogit.QlADr1,MELogit.QlADr17,MELogit.QlADr18,MELogit.QlADr19)
## Data: QLevelFactors
## Models:
## MELogit.QlADr1: AD ~ 1 + (1 | Participant)
## MELogit.QlADr17: AD ~ 1 + FirstBooklet + Interface + (1 | Participant)
## MELogit.QlADr19: AD ~ 1 + FirstBooklet * Interface + (1 | Participant)
## MELogit.QlADr18: AD ~ 1 + FirstBooklet + Interface + Order + (1 | Participant)
##           Df    AIC    BIC logLik deviance  Chisq Chi Df
## MELogit.QlADr1  2 1221.1 1231.5 -608.57  1217.1
## MELogit.QlADr17  4 1200.8 1221.6 -596.42  1192.8 24.3124    2
## MELogit.QlADr19  5 1202.7 1228.7 -596.36  1192.7  0.1146    1
## MELogit.QlADr18  7 1206.2 1242.5 -596.08  1192.2  0.5645    2
##           Pr(>Chisq)
## MELogit.QlADr1
## MELogit.QlADr17 5.256e-06 ***
## MELogit.QlADr19  0.7349
## MELogit.QlADr18  0.7541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model MEADr.glmer17 is the simplest with all main effects significant. A check shows the fit is not improved by additional random effects of TrueScenario:

```
anova(MELogit.QlADr17,MELogit.QlADr17.1)
## Data: QLevelFactors
## Models:
## MELogit.QlADr17: AD ~ 1 + FirstBooklet + Interface + (1 | Participant)
## MELogit.QlADr17.1: AD ~ 1 + FirstBooklet + Interface + (1 | Participant) + (1 |
## MELogit.QlADr17.1: TrueScenario)
##           Df    AIC    BIC logLik deviance  Chisq Chi Df
## MELogit.QlADr17  4 1200.8 1221.6 -596.42  1192.8
## MELogit.QlADr17.1 5 1202.6 1228.6 -596.31  1192.6 0.2221    1
##           Pr(>Chisq)
## MELogit.QlADr17
## MELogit.QlADr17.1  0.6374
```



**Figure 52 – Residual inspection plots for mixed-effects attempted diagnosis rate model MEADr.glmer17, Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.**

A rough estimate of the confidence intervals for the odds ratio of the main effect coefficients (estimated using standard deviation) are:

```
print(round(exp(melr.estimatedCI(MELogit.QIADr17)),2))
##           Est  LL  UL
## (Intercept)  3.20 2.10 4.86
## FirstBookletTRUE 1.58 1.17 2.13
## InterfaceC+R    1.86 1.38 2.51
```

A rough estimate of the model-described average marginal probability of marking a cause in the 2nd booklet with the C interface is:

```
print(round(probOfOdds(exp(fixef(MELogit.QIADr17)[1])),2))
## (Intercept)
##           0.76
```

A rough estimate of the average probability of marking a cause in the 1st booklet with C+R interface is:

```
print(round(probOfOdds(exp(sum(fixef(MELogit.QIADr17)))),2))
## [1] 0.90
```

However, interpreting probabilities is difficult without considering the random effects of Participant.

### D.4.3 Interface Effect on Diagnosis Accuracy Mixed Effect Logit Model

These models described the probability of a diagnosis being correct (RDr), given that a response had hit a change and the participant had attempted diagnosis (HD) (see Table 17 and Section

5.5.3.2). Again, a simple model (average probability only) was compared to increasing numbers of fixed effects.

```

MELogit.Q1RDr1 <- glmer(RightDiag_likely ~ (1 | TrueScenario)
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1RDr1, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: RightDiag_likely ~ (1 | TrueScenario)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##  608.0    616.3   -302.0   604.0     465
##
## Scaled residuals:
##   Min      1Q  Median      3Q      Max
## -0.766 -0.735 -0.713  1.339  1.409
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## TrueScenario (Intercept) 0.0153   0.124
## Number of obs: 467, groups: TrueScenario, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.613     0.109   -5.63  1.8e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

MELogit.Q1RDr19 <- glmer(RightDiag_likely ~ Interface + FirstBooklet + School + Order
+ (1 | TrueScenario)
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1RDr19, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: RightDiag_likely ~ Interface + FirstBooklet + School + Order +
## (1 | TrueScenario)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##  608.7    641.9   -296.4   592.7     459
##
## Scaled residuals:
##   Min      1Q  Median      3Q      Max
## -1.055 -0.765 -0.648  1.186  2.072
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## TrueScenario (Intercept) 0.0216   0.147

```

```

## Number of obs: 467, groups: TrueScenario, 10
##
## Fixed effects:
##      Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.047    0.251  -4.17   3e-05 ***
## InterfaceC+R    0.573    0.202   2.83  0.0047 **
## FirstBookletTRUE 0.222    0.202   1.10  0.2711
## Schools        0.174    0.229   0.76  0.4469
## OrderB         0.079    0.275   0.29  0.7737
## OrderC        -0.101    0.307  -0.33  0.7428
## OrderD        -0.300    0.297  -1.01  0.3136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation of fixed effects could have been required in summary()

##
## Correlation of Fixed Effects:
##      (Intr) IntC+R FBTRUE ScholS OrderB OrderC
## InterfacC+R -0.516
## FrstBklTRUE -0.354  0.085
## Schools     -0.190  0.039  0.018
## OrderB      -0.417 -0.030 -0.180 -0.086
## OrderC      -0.383  0.053 -0.085 -0.462  0.464
## OrderD      -0.389 -0.016 -0.146 -0.219  0.469  0.499

```

```

MELogit.Q1RDr18 <- glmer(RightDiag_likely ~ Interface + FirstBooklet + Order + (1 |
TrueScenario)
, data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1RDr18, corr = FALSE)

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula:
## RightDiag_likely ~ Interface + FirstBooklet + Order + (1 | TrueScenario)
## Data: QLevelFactors
##
##      AIC      BIC    logLik deviance df.resid
## 607.3    636.3   -296.7   593.3     460
##
## Scaled residuals:
##      Min      1Q  Median      3Q      Max
## -0.990 -0.772 -0.638  1.216  1.989
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## TrueScenario (Intercept) 0.022   0.148
## Number of obs: 467, groups: TrueScenario, 10
##
## Fixed effects:
##      Estimate Std. Error z value Pr(>|z|)

```

```

## (Intercept)      -1.01163    0.24633   -4.11    4e-05 ***
## InterfaceC+R     0.56721    0.20201    2.81    0.005 **
## FirstBookletTRUE 0.21985    0.20159    1.09    0.275
## OrderB           0.09673    0.27353    0.35    0.724
## OrderC           0.00693    0.27234    0.03    0.980
## OrderD          -0.25098    0.28971   -0.87    0.386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation of fixed effects could have been required in summary()

##
## Correlation of Fixed Effects:
##           (Intr) IntC+R FBTRUE OrderB OrderC
## InterfacC+R -0.518
## FrstBklTRUE -0.358  0.084
## OrderB      -0.442 -0.028 -0.179
## OrderC      -0.540  0.079 -0.086  0.479
## OrderD      -0.450 -0.007 -0.145  0.463  0.460

```

```

MELogit.Q1RDr17 <- glmer(RightDiag_likely ~ Interface + FirstBooklet + (1 |
TrueScenario)
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1RDr17, corr = FALSE)

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: RightDiag_likely ~ Interface + FirstBooklet + (1 | TrueScenario)
## Data: QLevelFactors
##
##           AIC          BIC    logLik deviance df.resid
##          602.9          619.4   -297.4    594.9     463
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.921 -0.773 -0.609  1.188  1.762
##
## Random effects:
## Groups          Name          Variance Std.Dev.
## TrueScenario (Intercept) 0.0177   0.133
## Number of obs: 467, groups: TrueScenario, 10
##
## Fixed effects:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.042     0.194   -5.37  7.8e-08 ***
## InterfaceC+R       0.570     0.200    2.85  0.0044 **
## FirstBookletTRUE   0.218     0.197    1.10  0.2692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

```

```
## Correlation of fixed effects could have been required in summary()

##
## Correlation of Fixed Effects:
##           (Intr) IntC+R
## InterfacC+R -0.624
## FrstBklTRUE -0.570  0.064
```

```
MELogit.QlRDr16 <- glmer(RightDiag_likely ~ Interface + (1 | TrueScenario)
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.QlRDr16, corr = FALSE)

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: RightDiag_likely ~ Interface + (1 | TrueScenario)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##  602.1    614.5   -298.0   596.1     464
##
## Scaled residuals:
##   Min     1Q  Median     3Q      Max
## -0.873 -0.795 -0.627  1.200  1.662
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## TrueScenario (Intercept) 0.0183  0.135
## Number of obs: 467, groups: TrueScenario, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.922     0.159   -5.79  7.2e-09 ***
## InterfacC+R    0.557     0.199    2.80  0.0052 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation of fixed effects could have been required in summary()

##
## Correlation of Fixed Effects:
##           (Intr)
## InterfacC+R -0.716
```

Comparing the least to most complex model options by likelihood ratio estimates, the model MELogit.QlRDr16 with Interface fixed effects is the most likely:

```
anova(MELogit.QlRDr1, MELogit.QlRDr16, MELogit.QlRDr17, MELogit.QlRDr18,
MELogit.QlRDr19)

## Data: QLevelFactors
```

```

## Models:
## MLogit.Q1RDr1: RightDiag_likely ~ (1 | TrueScenario)
## MLogit.Q1RDr16: RightDiag_likely ~ Interface + (1 | TrueScenario)
## MLogit.Q1RDr17: RightDiag_likely ~ Interface + FirstBooklet + (1 | TrueScenario)
## MLogit.Q1RDr18: RightDiag_likely ~ Interface + FirstBooklet + Order + (1 |
TrueScenario)
## MLogit.Q1RDr19: RightDiag_likely ~ Interface + FirstBooklet + School + Order +
## MLogit.Q1RDr19:      (1 | TrueScenario)
##           Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## MLogit.Q1RDr1   2 608 616   -302     604
## MLogit.Q1RDr16  3 602 615   -298     596  7.97     1  0.0048 **
## MLogit.Q1RDr17  4 603 619   -297     595  1.22     1  0.2686
## MLogit.Q1RDr18  7 607 636   -297     593  1.54     3  0.6724
## MLogit.Q1RDr19  8 609 642   -296     593  0.58     1  0.4470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Interface has a significant positive effect on likelihood of correctly diagnosing a detected hit. Random effects were forward-fit for the best model. No additional random effects were justified:

```

ranef(MLogit.Q1RDr16)

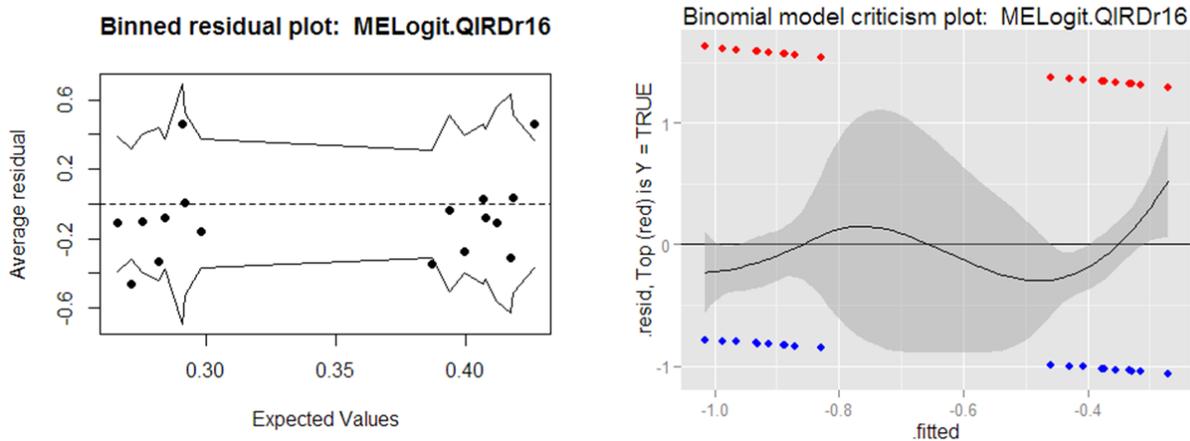
## $TrueScenario
##   (Intercept)
## 1      0.035918
## 2      0.010418...
...

anova(MLogit.Q1RDr16,MLogit.Q1RDr16.1,MLogit.Q1RDr16.2)

## Data: QLevelFactors
## Models:
## MLogit.Q1RDr16: RightDiag_likely ~ Interface + (1 | TrueScenario)
## MLogit.Q1RDr16.1: RightDiag_likely ~ Interface + (1 | Participant) + (1 |
TrueScenario)
## MLogit.Q1RDr16.2: RightDiag_likely ~ Interface + (1 | Participant:Order) + (1 |
## MLogit.Q1RDr16.2:      TrueScenario)
##           Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## MLogit.Q1RDr16   3 602 615   -298     596
## MLogit.Q1RDr16.1  4 604 621   -298     596  0     1  1
## MLogit.Q1RDr16.2  4 604 621   -298     596  0     0  1

```

Participant and Order do not account for much more variance and it is justified to use the simplest, TrueScenario-only random effect in the Question-level Right Diagnosis model #16.



**Figure 53 – Residual inspection plots for model MELogit.Q1RDr16, explaining likelihood of a hit with diagnosis being right. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.**

Rough estimates of the odds ratio confidence intervals (using the standard error of the coefficients) are:

```
print(round(exp(melr.estimatedCI(MELogit.Q1RDr16)),2))
##           Est  LL  UL
## (Intercept) 0.40 0.29 0.54
## InterfaceC+R 1.75 1.18 2.58
```

To corroborate the model fit, a rough estimate of average probability of correctly diagnosing a diagnosed hit with C interface:

```
print(round(probOfOdds(exp(fixef(MELogit.Q1RDr16)[1])),2)) # Intercept only
## (Intercept)
##           0.28
```

Same, with C+R interface (See Figure 37 for boxplots of observed data):

```
print(round(probOfOdds(exp(sum(fixef(MELogit.Q1RDr16))))),2))
## [1] 0.41
```

#### D.4.4 By-Interface Overall Performance Effect Mixed Effect Logit Models

This section describes the models used to quantify the likelihood of any given response being detected and correctly diagnosed, as discussed in Section E.1.3.

```
MELogit.Q1CR1 <- glmer(CR ~ (1 | TrueScenario)
                      , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1CR1, corr = FALSE)
```

```

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: CR ~ (1 | TrueScenario)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##  992.1  1002.5  -494.0   988.1   1327
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.4114 -0.3838 -0.3704 -0.3348  2.9865
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## TrueScenario (Intercept) 0.03948  0.1987
## Number of obs: 1329, groups: TrueScenario, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9612     0.1052  -18.64  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

MELogit.Q1CR19 <- glmer(CR ~ Interface + FirstBooklet + School + Order + (1 |
TrueScenario)
                    , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1CR19, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula:
## CR ~ Interface + FirstBooklet + School + Order + (1 | TrueScenario)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##  988.9  1030.5  -486.5   972.9   1321
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.5106 -0.4214 -0.3473 -0.2871  3.8811
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## TrueScenario (Intercept) 0.04622  0.215
## Number of obs: 1329, groups: TrueScenario, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.358619   0.220581 -10.693  < 2e-16 ***
## InterfaceC+R   0.616245   0.174794  3.526  0.000423 ***
## FirstBookletTRUE 0.171671   0.174449  0.984  0.325079
## Schools        0.110196   0.197533  0.558  0.576940

```

```

## OrderB          -0.078351   0.233167  -0.336  0.736847
## OrderC          0.009856   0.264319   0.037  0.970256
## OrderD          -0.280917   0.262842  -1.069  0.285174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##          (Intr) IntC+R FBTRUE ScholS OrderB OrderC
## InterfacC+R -0.504
## FrstBklTRUE -0.352  0.093
## Schools     -0.142 -0.012 -0.020
## OrderB      -0.389 -0.068 -0.211 -0.070
## OrderC      -0.383  0.050 -0.041 -0.467  0.455
## OrderD      -0.343 -0.033 -0.155 -0.277  0.470  0.512

```

```

MELogit.Q1CR18 <- glmer(CR ~ Interface + FirstBooklet + Order + (1 | TrueScenario)
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1CR18, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: CR ~ Interface + FirstBooklet + Order + (1 | TrueScenario)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##  987.2  1023.6  -486.6   973.2   1322
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.4902 -0.4180 -0.3507 -0.2784  3.7748
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## TrueScenario (Intercept) 0.04611  0.2147
## Number of obs: 1329, groups: TrueScenario, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.34190    0.21831 -10.727 < 2e-16 ***
## InterfaceC+R    0.61757    0.17476   3.534 0.000409 ***
## FirstBookletTRUE 0.17363    0.17440   0.996 0.319460
## OrderB        -0.06951    0.23260  -0.299 0.765051
## OrderC         0.07865    0.23370   0.337 0.736460
## OrderD        -0.24087    0.25251  -0.954 0.340129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##          (Intr) IntC+R FBTRUE OrderB OrderC

```

```
## InterfacC+R -0.511
## FrstBklTRUE -0.358 0.093
## OrderB      -0.404 -0.069 -0.214
## OrderC      -0.514 0.051 -0.056 0.479
## OrderD      -0.402 -0.038 -0.166 0.470 0.450
```

```
MELogit.Q1CR17 <- glmer(CR ~ Interface + FirstBooklet + (1 | TrueScenario)
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1CR17, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: CR ~ Interface + FirstBooklet + (1 | TrueScenario)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##  982.9  1003.7  -487.5   974.9   1325
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.4887 -0.4199 -0.3372 -0.2998  3.5520
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## TrueScenario (Intercept) 0.04415  0.2101
## Number of obs: 1329, groups: TrueScenario, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.3670    0.1756 -13.481 < 2e-16 ***
## InterfaceC+R    0.6018    0.1727  3.484 0.000493 ***
## FirstBookletTRUE 0.1406    0.1687  0.833 0.404719
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##              (Intr) IntC+R
## InterfacC+R -0.599
## FrstBklTRUE -0.516 0.003
```

```
MELogit.Q1CR16 <- glmer(CR ~ Interface + (1 | TrueScenario)
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.Q1CR16, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: CR ~ Interface + (1 | TrueScenario)
## Data: QLevelFactors
##
```

```

##      AIC      BIC  logLik deviance df.resid
##    981.6    997.2  -487.8   975.6    1326
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.4736 -0.4202 -0.3362 -0.3065  3.5659
##
## Random effects:
##  Groups      Name      Variance Std.Dev.
## TrueScenario (Intercept) 0.04484  0.2118
## Number of obs: 1329, groups: TrueScenario, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.2935     0.1506 -15.230 < 2e-16 ***
## InterfaceC+R   0.6017     0.1727   3.485 0.000492 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

```

Comparing the alternate models:

```

anova(MELogit.Q1CR1, MELogit.Q1CR16, MELogit.Q1CR17, MELogit.Q1CR18, MELogit.Q1CR19)
## Data: QLevelFactors
## Models:
## MELogit.Q1CR1: CR ~ (1 | TrueScenario)
## MELogit.Q1CR16: CR ~ Interface + (1 | TrueScenario)
## MELogit.Q1CR17: CR ~ Interface + FirstBooklet + (1 | TrueScenario)
## MELogit.Q1CR18: CR ~ Interface + FirstBooklet + Order + (1 | TrueScenario)
## MELogit.Q1CR19: CR ~ Interface + FirstBooklet + School + Order + (1 |
TrueScenario)
##              Df      AIC      BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## MELogit.Q1CR1  2 992.09 1002.48 -494.05  988.09
## MELogit.Q1CR16  3 981.62  997.19 -487.81  975.62 12.4788      1 0.0004116 ***
## MELogit.Q1CR17  4 982.92 1003.69 -487.46  974.92  0.6938      1 0.4048718
## MELogit.Q1CR18  7 987.23 1023.57 -486.61  973.23  1.6936      3 0.6383599
## MELogit.Q1CR19  8 988.92 1030.46 -486.46  972.92  0.3093      1 0.5781225
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

```

So the simplest model with Interface as main effect is the most likely, with lowest AIC and BIC.

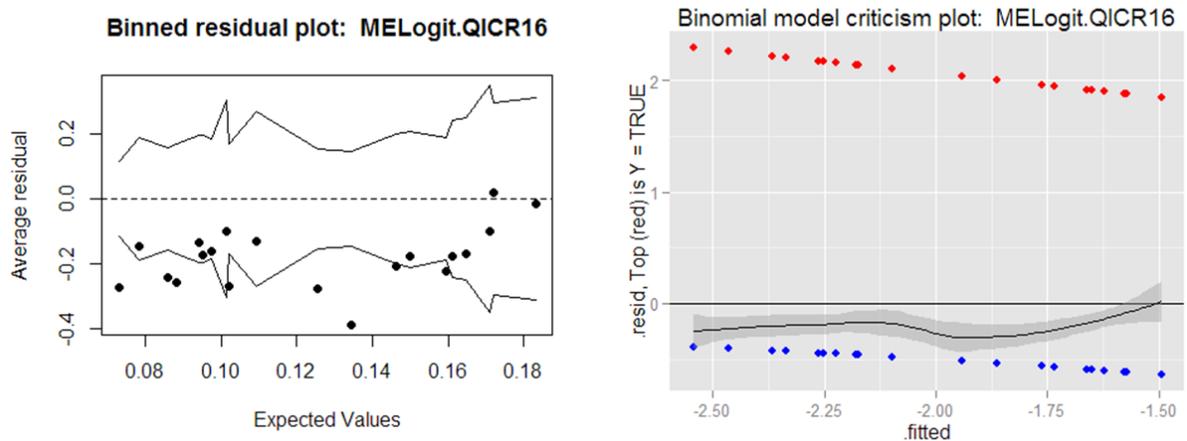
Next, check whether the TrueScenario random effect is sufficient using a comparison model:

```

anova(MELogit.Q1CR16, MELogit.Q1CR16.1)
## Data: QLevelFactors
## Models:
## MELogit.Q1CR16: CR ~ Interface + (1 | TrueScenario)
## MELogit.Q1CR16.1: CR ~ Interface + (1 | TrueScenario) + (1 | Participant)
##              Df      AIC      BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## MELogit.Q1CR16  3 981.62  997.19 -487.81  975.62
## MELogit.Q1CR16.1  4 983.62 1004.38 -487.81  975.62      0      1      1
##

```

Again, participant accounts for no significant additional random variance.



**Figure 54 – Residual inspection plots for model MELogit.Q1CR16, explaining likelihood of a response being completely correct. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.**

Using the best model, what are the odds ratios for the main effect of Interface?

```
print(round(exp(melr.estimatedCI(MELogit.Q1CR16)),2))
##           Est  LL  UL
## (Intercept) 0.10 0.08 0.14
## InterfaceC+R 1.83 1.30 2.56
```

A rough estimate of average probability of a response being a hit with correct diagnosis with C interface:

```
print(round(probOfOdds(exp(fixef(MELogit.Q1CR16)[1])),2)) # Intercept only
## (Intercept)
##           0.09
```

A rough estimate of average probability of a response being a hit with correct diagnosis with C+R interface (consistent with Figure 37):

```
print(round(probOfOdds(exp(sum(fixef(MELogit.Q1CR16)))),2))
## [1] 0.16
```

## D.5 By-Scenario Performance

### D.5.1 Scenario Effects on Detection

```
describeBy(TLevelFactors, TLevelFactors$Scenario)
## group: 1
```

```

##          vars  n  mean  sd median trimmed  mad min max range
## TrueScenario*    1 66  3.50 2.52   3.5   3.50  3.71  1  6   5
## Participant*     2 66 17.00 9.59  17.0  17.00 11.86  1 33  32
## Order*           3 66  2.45 1.14   2.0   2.44  1.48  1  4   3
## FirstBooklet*    4 66  0.50 0.50   0.5   0.50  0.74  0  1   1
## Interface*       5 66  1.50 0.50   1.5   1.50  0.74  1  2   1
## Scenario*        6 66  1.00 0.00   1.0   1.00  0.00  1  1   0
## Change_inbox     7 66  0.95 0.83   1.0   0.94  1.48  0  2   2
## Diag_inbox       8 66  0.74 0.73   1.0   0.69  1.48  0  2   2
## RightDiag_inbox  9 66  0.21 0.45   0.0   0.13  0.00  0  2   2
## Change_likely    10 66  1.65 0.62   2.0   1.78  0.00  0  2   2
## Diag_likely      11 66  1.35 0.73   1.5   1.43  0.74  0  2   2
## RightDiag_likely 12 66  0.50 0.56   0.0   0.46  0.00  0  2   2
## MarkedCause      13 66  3.23 1.57   3.0   3.17  1.48  0  8   8
## Responses        14 66  4.02 1.49   4.0   3.83  1.48  2  9   7
## TrueChanges      15 66  2.00 0.00   2.0   2.00  0.00  2  2   0
## -----
## group: 2
##          vars  n  mean  sd median trimmed  mad min max range
## TrueScenario*    1 66  4.50 2.52   4.5   4.50  3.71  2  7   5
## Participant*     2 66 17.00 9.59  17.0  17.00 11.86  1 33  32
## Order*           3 66  2.45 1.14   2.0   2.44  1.48  1  4   3
## FirstBooklet*    4 66  0.50 0.50   0.5   0.50  0.74  0  1   1
## Interface*       5 66  1.50 0.50   1.5   1.50  0.74  1  2   1
## Scenario*        6 66  2.00 0.00   2.0   2.00  0.00  2  2   0
## Change_inbox     7 66  1.41 0.70   1.0   1.39  0.00  0  3   3
## Diag_inbox       8 66  1.06 0.70   1.0   1.07  0.00  0  2   2
## RightDiag_inbox  9 66  0.36 0.57   0.0   0.28  0.00  0  2   2
## Change_likely    10 66  1.71 0.82   2.0   1.67  1.48  0  3   3
## Diag_likely      11 66  1.27 0.83   1.0   1.26  1.48  0  3   3
## RightDiag_likely 12 66  0.45 0.66   0.0   0.35  0.00  0  3   3
## MarkedCause      13 66  1.88 1.34   2.0   1.80  1.48  0  5   5
## Responses        14 66  2.45 1.39   2.0   2.31  1.48  1  6   5
## TrueChanges      15 66  3.00 0.00   3.0   3.00  0.00  3  3   0
## -----
## group: 3
##          vars  n  mean  sd median trimmed  mad min max range
## TrueScenario*    1 66  5.50 2.52   5.5   5.50  3.71  3  8   5
## Participant*     2 66 17.00 9.59  17.0  17.00 11.86  1 33  32
## Order*           3 66  2.45 1.14   2.0   2.44  1.48  1  4   3
## FirstBooklet*    4 66  0.50 0.50   0.5   0.50  0.74  0  1   1
## Interface*       5 66  1.50 0.50   1.5   1.50  0.74  1  2   1
## Scenario*        6 66  3.00 0.00   3.0   3.00  0.00  3  3   0
## Change_inbox     7 66  1.58 0.70   2.0   1.63  0.00  0  3   3
## Diag_inbox       8 66  1.32 0.75   1.0   1.35  1.48  0  3   3
## RightDiag_inbox  9 66  0.47 0.61   0.0   0.39  0.00  0  2   2
## Change_likely    10 66  1.80 0.71   2.0   1.80  0.00  0  3   3
## Diag_likely      11 66  1.45 0.81   1.5   1.46  0.74  0  3   3
## RightDiag_likely 12 66  0.48 0.61   0.0   0.41  0.00  0  2   2
## MarkedCause      13 66  3.68 2.12   3.0   3.65  1.48  0  9   9
## Responses        14 66  4.67 2.14   4.0   4.56  2.22  1 10   9
## TrueChanges      15 66  3.00 0.00   3.0   3.00  0.00  3  3   0
## -----
## group: 4
##          vars  n  mean  sd median trimmed  mad min max range
## TrueScenario*    1 66  6.50 2.52   6.5   6.50  3.71  4  9   5

```

```

## Participant*      2 66 17.00 9.59   17.0   17.00 11.86   1 33 32
## Order*           3 66  2.45 1.14    2.0    2.44  1.48   1  4  3
## FirstBooklet*   4 66  0.50 0.50    0.5    0.50  0.74   0  1  1
## Interface*      5 66  1.50 0.50    1.5    1.50  0.74   1  2  1
## Scenario*       6 66  4.00 0.00    4.0    4.00  0.00   4  4  0
## Change_inbox    7 66  1.42 0.98    1.0    1.37  1.48   0  4  4
## Diag_inbox      8 66  1.24 1.01    1.0    1.15  1.48   0  4  4
## RightDiag_inbox 9 66  0.26 0.54    0.0    0.15  0.00   0  2  2
## Change_likely   10 66  2.59 0.82    3.0    2.59  1.48   1  4  3
## Diag_likely     11 66  2.17 1.00    2.0    2.15  1.48   0  4  4
## RightDiag_likely 12 66  0.62 0.76    0.5    0.50  0.74   0  4  4
## MarkedCause     13 66  3.79 1.79    4.0    3.74  1.48   0  8  8
## Responses       14 66  4.61 1.69    4.0    4.52  1.48   2  9  7
## TrueChanges     15 66  4.00 0.00    4.0    4.00  0.00   4  4  0
## -----
## group: 5
##          vars  n  mean  sd median trimmed  mad min max range
## TrueScenario*   1 66  7.50 2.52   7.5   7.50  3.71   5 10   5
## Participant*   2 66 17.00 9.59  17.0  17.00 11.86   1 33  32
## Order*         3 66  2.45 1.14   2.0   2.44  1.48   1  4   3
## FirstBooklet*  4 66  0.50 0.50   0.5   0.50  0.74   0  1   1
## Interface*     5 66  1.50 0.50   1.5   1.50  0.74   1  2   1
## Scenario*      6 66  5.00 0.00   5.0   5.00  0.00   5  5   0
## Change_inbox   7 66  0.85 0.36   1.0   0.93  0.00   0  1   1
## Diag_inbox     8 66  0.73 0.45   1.0   0.78  0.00   0  1   1
## RightDiag_inbox 9 66  0.35 0.48   0.0   0.31  0.00   0  1   1
## Change_likely  10 66  0.97 0.17   1.0   1.00  0.00   0  1   1
## Diag_likely    11 66  0.83 0.38   1.0   0.91  0.00   0  1   1
## RightDiag_likely 12 66  0.41 0.50   0.0   0.39  0.00   0  1   1
## MarkedCause    13 66  3.65 1.58   3.5   3.70  2.22   0  6   6
## Responses      14 66  4.39 1.54   5.0   4.48  1.48   1  7   6
## TrueChanges    15 66  1.00 0.00   1.0   1.00  0.00   1  1   0

```

## D.5.2 Scenario Effect on Diagnosis Mixed Effect Logit Models

This mixed effect generalized binomial model was developed to explain Right Diagnosis Rate (RDr) with fixed effects of Scenario.

```

# Use a binomial model. Again, RightDiag_likely is only available for cases where a
# response has
# been matched to a change and participant has chosen to diagnose. Do not need
# offset/exposure.

MELogit.TfRDr.1 <- glmer(RightDiag_likely ~ Scenario + (1 | Participant )
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.TfRDr.1, corr = FALSE)

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: RightDiag_likely ~ Scenario + (1 | Participant)
## Data: QLevelFactors
##

```

```

##      AIC      BIC  logLik deviance df.resid
##    608.7    633.5  -298.3   596.7     461
##
## Scaled residuals:
##   Min      1Q  Median      3Q      Max
## -0.982 -0.745 -0.634  1.303  1.577
##
## Random effects:
##   Groups      Name      Variance Std.Dev.
## Participant (Intercept) 0          0
## Number of obs: 467, groups: Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.5288    0.2195  -2.41   0.016 *
## Scenario2    -0.0589    0.3162  -0.19   0.852
## Scenario3    -0.1643    0.3083  -0.53   0.594
## Scenario4    -0.3826    0.2870  -1.33   0.183
## Scenario5     0.4925    0.3477   1.42   0.157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

MELogit.TfRDr.19 <- glmer(RightDiag_likely ~ Scenario * Interface + (1 | Participant
)
, data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.TfRDr.19, corr = FALSE)

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: RightDiag_likely ~ Scenario * Interface + (1 | Participant)
## Data: QLevelFactors
##
##      AIC      BIC  logLik deviance df.resid
##    602.5    648.2  -290.3   580.5     456
##
## Scaled residuals:
##   Min      1Q  Median      3Q      Max
## -1.035 -0.663 -0.616  1.080  2.160
##
## Random effects:
##   Groups      Name      Variance Std.Dev.
## Participant (Intercept) 0          0
## Number of obs: 467, groups: Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.9694    0.3541  -2.74   0.0062 **
## Scenario2     0.1484    0.5063   0.29   0.7694
## Scenario3    -0.5710    0.5102  -1.12   0.2630
## Scenario4     0.0531    0.4507   0.12   0.9062
## Scenario5     0.8152    0.5293   1.54   0.1235
## InterfaceC+R  0.7646    0.4559   1.68   0.0935 .

```

```
## Scenario2:InterfaceC+R -0.3665    0.6526  -0.56   0.5744
## Scenario3:InterfaceC+R  0.8203    0.6571   1.25   0.2119
## Scenario4:InterfaceC+R -0.7559    0.5888  -1.28   0.1992
## Scenario5:InterfaceC+R -0.5415    0.7076  -0.76   0.4442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MELogit.TfRDr.18 <- glmer(RightDiag_likely ~ Scenario + Interface + (1 | Participant
)
, data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.TfRDr.18, corr = FALSE)

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: RightDiag_likely ~ Scenario + Interface + (1 | Participant)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##  602.4    631.5   -294.2   588.4     460
##
## Scaled residuals:
##   Min      1Q  Median      3Q      Max
## -1.123 -0.713 -0.614  1.184  1.867
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## Participant (Intercept) 1.65e-14 1.29e-07
## Number of obs: 467, groups: Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.8541    0.2511  -3.40  0.00067 ***
## Scenario2     -0.0719    0.3192  -0.23  0.82167
## Scenario3     -0.1202    0.3115  -0.39  0.69950
## Scenario4     -0.3942    0.2895  -1.36  0.17333
## Scenario5      0.5156    0.3513   1.47  0.14220
## InterfaceC+R  0.5712    0.2008   2.84  0.00445 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

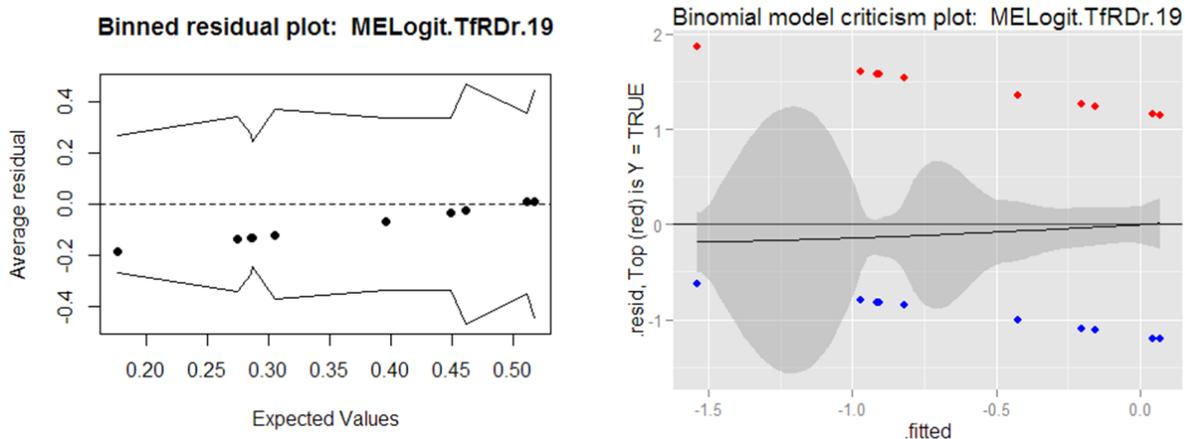
Comparing goodness-of-fit of the simplest (Scenario only) to most complicated (Scenario and Interface interaction effect):

```
anova(MELogit.TfRDr.1, MELogit.TfRDr.18, MELogit.TfRDr.19)

## Data: QLevelFactors
## Models:
## MELogit.TfRDr.1: RightDiag_likely ~ Scenario + (1 | Participant)
## MELogit.TfRDr.18: RightDiag_likely ~ Scenario + Interface + (1 | Participant)
## MELogit.TfRDr.19: RightDiag_likely ~ Scenario * Interface + (1 | Participant)
```

```
##           Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## MLogit.TfRDr.1   6 609 634  -298     597
## MLogit.TfRDr.18  7 602 631  -294     588  8.24    1  0.0041 **
## MLogit.TfRDr.19 11 603 648  -290     581  7.89    4  0.0958 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the viewpoint of Scenario as a fixed factor, Interface is still significant. It is arguable whether a Scenario \* Interface interaction is significant globally. However, these models may still be useful for contrasting Scenarios against each other.



**Figure 55 – Residual inspection plots for model MLogit.TfRDr.19, explaining likelihood of a hit diagnosis being right. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.**

Estimated Confidence Intervals and consistency check for probability of Right Diagnosis in 1<sup>st</sup> Scenario C condition:

```
print(round(exp(melr.estimatedCI(MLogit.TfRDr.19)),2))

##           Est  LL  UL
## (Intercept)  0.38 0.19 0.76
## Scenario2    1.16 0.43 3.13
## Scenario3    0.56 0.21 1.54
## Scenario4    1.05 0.44 2.55
## Scenario5    2.26 0.80 6.38
## InterfaceC+R  2.15 0.88 5.25
## Scenario2:InterfaceC+R 0.69 0.19 2.49
## Scenario3:InterfaceC+R 2.27 0.63 8.23
## Scenario4:InterfaceC+R 0.47 0.15 1.49
## Scenario5:InterfaceC+R 0.58 0.15 2.33

print(round(probOfOdds(exp(fixef(MLogit.TfRDr.19)[1])),2))

## (Intercept)
```

```
## 0.28
```

### D.5.3 Contrasting Scenario Diagnosis Performance

In the C only interface condition, diagnosis performance in Scenario 5 seems higher than average. Using the Scenario \* Interface interaction model described above, a contrast demonstrated statistically significant differences between levels (though uncorrected for familywise error):

```
MELogit.TfRDr.19.c1 <- rbind(
"C, Scenario 1 vs. average" = c(0,-0.2,-0.2,-0.2,-0.2,0,0,0,0,0),
"C, Scenario 2 vs. average" = c(0,0.8,-0.2,-0.2,-0.2,0,0,0,0,0),
"C, Scenario 3 vs. average" = c(0,-0.2,0.8,-0.2,-0.2,0,0,0,0,0),
"C, Scenario 4 vs. average" = c(0,-0.2,-0.2,0.8,-0.2,0,0,0,0,0),
"C, Scenario 5 vs. average" = c(0,-0.2,-0.2,-0.2,0.8,0,0,0,0,0)
)

summary(glht(MELogit.TfRDr.19, MELogit.TfRDr.19.c1), test=adjusted("none"))

##
## Simultaneous Tests for General Linear Hypotheses
##
## Fit: glmer(formula = RightDiag_likely ~ Scenario * Interface + (1 |
## Participant), data = QLevelFactors, family = binomial, nAGQ = 1)
##
## Linear Hypotheses:
##
## Estimate Std. Error z value Pr(>|z|)
## C, Scenario 1 vs. average == 0 -0.0891 0.3165 -0.28 0.778
## C, Scenario 2 vs. average == 0 0.0593 0.3217 0.18 0.854
## C, Scenario 3 vs. average == 0 -0.6602 0.3254 -2.03 0.042 *
## C, Scenario 4 vs. average == 0 -0.0360 0.2676 -0.13 0.893
## C, Scenario 5 vs. average == 0 0.7261 0.3432 2.12 0.034 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

Similarly, in the C+R condition, Scenario 4 diagnosis performance is significantly lower than the average of the other four scenarios.

```
MELogit.TfRDr.19.c3 <- rbind(
"C+R, Scenario 1 vs. average" = c(0,-0.2,-0.2,-0.2,-0.2,0,-0.2,-0.2,-0.2,-0.2),
"C+R, Scenario 2 vs. average" = c(0,0.8,-0.2,-0.2,-0.2,0,0.8,-0.2,-0.2,-0.2),
"C+R, Scenario 3 vs. average" = c(0,-0.2,0.8,-0.2,-0.2,0,-0.2,0.8,-0.2,-0.2),
"C+R, Scenario 4 vs. average" = c(0,-0.2,-0.2,0.8,-0.2,0,-0.2,-0.2,0.8,-0.2),
"C+R, Scenario 5 vs. average" = c(0,-0.2,-0.2,-0.2,0.8,0,-0.2,-0.2,-0.2,0.8)
)

summary(glht(MELogit.TfRDr.19, MELogit.TfRDr.19.c3), test=adjusted("none"))

##
## Simultaneous Tests for General Linear Hypotheses
```

```
##
## Fit: glmer(formula = RightDiag_likely ~ Scenario * Interface + (1 |
##   Participant), data = QLevelFactors, family = binomial, nAGQ = 1)
##
## Linear Hypotheses:
##
##           Estimate Std. Error z value Pr(>|z|)
## C+R, Scenario 1 vs. average == 0  0.0796    0.2604   0.31  0.7600
## C+R, Scenario 2 vs. average == 0 -0.1385    0.2657  -0.52  0.6021
## C+R, Scenario 3 vs. average == 0  0.3288    0.2677   1.23  0.2194
## C+R, Scenario 4 vs. average == 0 -0.6232    0.2343  -2.66  0.0078 **
## C+R, Scenario 5 vs. average == 0  0.3533    0.3181   1.11  0.2666
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

The model suggests that practical differences between diagnosis performance in C and C+R interface conditions are mainly apparent in Scenarios 1 and 3:

```
> MELogit.TfRDr.19.c5 <- rbind("Scenario 1, C vs. C+R" = c(0,0,0,0,0,1,0,0,0,0),
+                             "Scenario 2, C vs. C+R" = c(0,0,0,0,0,1,1,0,0,0),
+                             "Scenario 3, C vs. C+R" = c(0,0,0,0,0,1,0,1,0,0),
+                             "Scenario 4, C vs. C+R" = c(0,0,0,0,0,1,0,0,1,0),
+                             "Scenario 5, C vs. C+R" = c(0,0,0,0,0,1,0,0,0,1))

summary(glht(MELogit.TfRDr.19, MELogit.TfRDr.19.c5), test=adjusted("none"))

##
## Simultaneous Tests for General Linear Hypotheses
##
## Fit: glmer(formula = RightDiag_likely ~ Scenario * Interface + (1 |
##   Participant), data = QLevelFactors, family = binomial, nAGQ = 1)
##
## Linear Hypotheses:
##
##           Estimate Std. Error z value Pr(>|z|)
## Scenario 1, C vs. C+R == 0  0.76461    0.45593   1.68  0.09354 .
## Scenario 2, C vs. C+R == 0  0.39812    0.46693   0.85  0.39385
## Scenario 3, C vs. C+R == 0  1.58490    0.47313   3.35  0.00081 ***
## Scenario 4, C vs. C+R == 0  0.00873    0.37256   0.02  0.98130
## Scenario 5, C vs. C+R == 0  0.22314    0.54116   0.41  0.68009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

#### D.5.4 Scenario Performance Difference Mixed Effect Logit Models

The mixed effects logit regressions used for the analysis described in E.1.3 were formulated similarly as in D.4.4 above.

In the interests of brevity, only the null model, the likelihood ratio tests for the alternative models, and the two best models are presented here:

```

MELogit.TfCRr.1 <- glmer(CR ~ Scenario + (1 | Participant )
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.TfCRr.1, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: CR ~ Scenario + (1 | Participant)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##    992.0   1023.2   -490.0   980.0    1323
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.4767 -0.3948 -0.3405 -0.3204  3.1210
##
## Random effects:
## Groups      Name                Variance Std.Dev.
## Participant (Intercept) 0          0
## Number of obs: 1329, groups: Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.95023    0.18604  -10.483  <2e-16 ***
## Scenario2    0.46863    0.27481   1.705   0.0881 .
## Scenario3   -0.20444    0.26360  -0.776   0.4380
## Scenario4    0.09165    0.25060   0.366   0.7146
## Scenario5   -0.32609    0.27468  -1.187   0.2352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##              (Intr) Scenr2 Scenr3 Scenr4
## Scenario2 -0.677
## Scenario3 -0.706  0.478
## Scenario4 -0.742  0.503  0.524
## Scenario5 -0.677  0.459  0.478  0.503

```

```

MELogit.TfCRr.17 <- glmer(CR ~ Scenario * Interface + (1 | Participant )
                        , data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.TfCRr.17, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: CR ~ Scenario * Interface + (1 | Participant)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##    986.8   1043.9   -482.4   964.8    1318
##
## Scaled residuals:

```

```

##      Min      1Q  Median      3Q      Max
## -0.5492 -0.4256 -0.3384 -0.3015  4.0689
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## Participant (Intercept) 0          0
## Number of obs: 1329, groups: Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.36428   0.31533  -7.498 6.49e-14 ***
## Scenario2      0.52807   0.45259   1.167  0.2433
## Scenario3     -0.44244   0.46611  -0.949  0.3425
## Scenario4      0.32018   0.40270   0.795  0.4266
## Scenario5     -0.03362   0.43627  -0.077  0.9386
## InterfaceC+R   0.71039   0.39189   1.813  0.0699 .
## Scenario2:InterfaceC+R -0.07287   0.57227  -0.127  0.8987
## Scenario3:InterfaceC+R  0.38764   0.56811   0.682  0.4950
## Scenario4:InterfaceC+R -0.35107   0.51754  -0.678  0.4975
## Scenario5:InterfaceC+R -0.47964   0.56459  -0.850  0.3956
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##      (Intr) Scenr2 Scenr3 Scenr4 Scenr5 IntC+R S2:IC+ S3:IC+ S4:IC+
## Scenario2  -0.697
## Scenario3  -0.677  0.471
## Scenario4  -0.783  0.546  0.530
## Scenario5  -0.723  0.504  0.489  0.566
## InterfacC+R -0.805  0.561  0.544  0.630  0.582
## Scnr2:InC+R  0.551 -0.791 -0.373 -0.431 -0.398 -0.685
## Scnr3:InC+R  0.555 -0.387 -0.820 -0.435 -0.401 -0.690  0.472
## Scnr4:InC+R  0.609 -0.424 -0.412 -0.778 -0.440 -0.757  0.519  0.522
## Scnr5:InC+R  0.558 -0.389 -0.378 -0.437 -0.773 -0.694  0.475  0.479  0.526

```

```

MELogit.TfCRr.16 <- glmer(CR ~ Scenario + Interface + (1 | Participant )
, data = QLevelFactors, family = binomial, nAGQ = 1)
summary(MELogit.TfCRr.16, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: CR ~ Scenario + Interface + (1 | Participant)
## Data: QLevelFactors
##
##      AIC      BIC   logLik deviance df.resid
##   981.8  1018.1  -483.9   967.8    1322
##
## Scaled residuals:
##      Min      1Q  Median      3Q      Max
## -0.5447 -0.4043 -0.3363 -0.2894  3.6920
##

```

```

## Random effects:
## Groups      Name          Variance Std.Dev.
## Participant (Intercept) 2.848e-14 1.687e-07
## Number of obs: 1329, groups: Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.2915    0.2160 -10.606 < 2e-16 ***
## Scenario2     0.4802    0.2764  1.737 0.082334 .
## Scenario3    -0.1886    0.2648 -0.712 0.476351
## Scenario4     0.1123    0.2519  0.446 0.655801
## Scenario5    -0.3208    0.2758 -1.163 0.244720
## InterfaceC+R  0.5962    0.1728  3.450 0.000561 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##              (Intr) Scenr2 Scenr3 Scenr4 Scenr5
## Scenario2    -0.594
## Scenario3    -0.618  0.478
## Scenario4    -0.654  0.502  0.524
## Scenario5    -0.588  0.458  0.478  0.503
## InterfacC+R -0.502  0.017  0.015  0.024  0.003

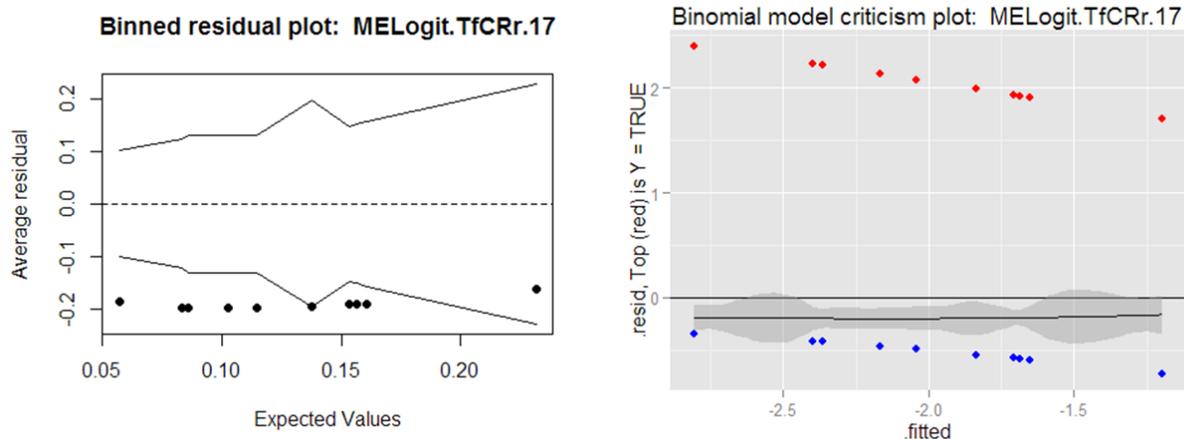
```

```

anova(MELogit.TfCRr.1, MELogit.TfCRr.16, MELogit.TfCRr.17, MELogit.TfCRr.18,
MELogit.TfCRr.19)
## Data: QLevelFactors
## Models:
## MELogit.TfCRr.1: CR ~ Scenario + (1 | Participant)
## MELogit.TfCRr.16: CR ~ Scenario + Interface + (1 | Participant)
## MELogit.TfCRr.17: CR ~ Scenario * Interface + (1 | Participant)
## MELogit.TfCRr.18: CR ~ Scenario * Interface + FirstBooklet + (1 | Participant)
## MELogit.TfCRr.19: CR ~ Scenario * Interface + FirstBooklet + School + (1 |
Participant)
##              Df      AIC      BIC  loglik deviance  Chisq Chi Df Pr(>Chisq)
## MELogit.TfCRr.1  6 992.01 1023.2 -490.01  980.01
## MELogit.TfCRr.16  7 981.76 1018.1 -483.88  967.76 12.2531      1 0.0004645 ***
## MELogit.TfCRr.17 11 986.81 1043.9 -482.40  964.81  2.9508      4 0.5660986
## MELogit.TfCRr.18 12 988.22 1050.5 -482.11  964.22  0.5929      1 0.4412930
## MELogit.TfCRr.19 13 989.76 1057.3 -481.88  963.76  0.4562      1 0.4994204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Despite the Scenario main and interaction effects not being significant in model MELogit.TfCRr.17, the experimental structure justifies including Scenario effects.



**Figure 56 – Residual Inspection plots for model MELogit.TfCRr.17, explaining likelihood of a response being completely correct. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.**

The Interface\*Scenario interaction effects model was used for contrasts on the different scenario levels, using the same contrasts as in D.5.3. In the C condition, comparing each scenario to an average, all (unadjusted)  $p > .1$ :

```
summary(glht(MELogit.TfCRr.17, MELogit.TfRDr.19.c1), test=adjusted("none"))
##
## Simultaneous Tests for General Linear Hypotheses
##
## Fit: glmer(formula = CR ~ Scenario * Interface + (1 | Participant),
## data = QLevelFactors, family = binomial, nAGQ = 1)
##
## Linear Hypotheses:
##              Estimate Std. Error z value Pr(>|z|)
## C, Scenario 1 vs. average == 0 -0.07444  0.28055  -0.265  0.7908
## C, Scenario 2 vs. average == 0  0.45363  0.28687  1.581  0.1138
## C, Scenario 3 vs. average == 0 -0.51688  0.29958  -1.725  0.0845 .
## C, Scenario 4 vs. average == 0  0.24574  0.23811  1.032  0.3021
## C, Scenario 5 vs. average == 0 -0.10805  0.27128  -0.398  0.6904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

In the C+R condition, Scenario 2 and Scenario 5 may have significantly better and worse overall performance, unadjusted  $p < .05$ :

```
summary(glht(MELogit.TfCRr.17, MELogit.TfRDr.19.c3), test=adjusted("none"))
##
## Simultaneous Tests for General Linear Hypotheses
##
## Fit: glmer(formula = CR ~ Scenario * Interface + (1 | Participant),
## data = QLevelFactors, family = binomial, nAGQ = 1)
##
##
```

```
## Linear Hypotheses:
##
##           Estimate Std. Error z value Pr(>|z|)
## C+R, Scenario 1 vs. average == 0  0.028753  0.210909  0.136  0.8916
## C+R, Scenario 2 vs. average == 0  0.483947  0.230428  2.100  0.0357 *
## C+R, Scenario 3 vs. average == 0 -0.026050  0.206895 -0.126  0.8998
## C+R, Scenario 4 vs. average == 0 -0.002145  0.207169 -0.010  0.9917
## C+R, Scenario 5 vs. average == 0 -0.484504  0.237850 -2.037  0.0416 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

Consistent with diagnosis performance, Scenarios 2,4 and 5 had the least evidence for benefits of the C+R interface condition,  $p > .12$ .

```
summary(glht(MELogit.TfCRr.17, MELogit.TfRDr.19.c5), test=adjusted("none"))

##
## Simultaneous Tests for General Linear Hypotheses
##
## Fit: glmer(formula = isTRUE(RightDiag_likely) ~ Scenario * Interface +
## (1 | Participant), data = QLevelFactors, family = binomial,
## nAGQ = 1)
##
## Linear Hypotheses:
##           Estimate Std. Error z value Pr(>|z|)
## Scenario 1, C vs. C+R == 0  0.710  0.392  1.81  0.0699 .
## Scenario 2, C vs. C+R == 0  0.638  0.417  1.53  0.1263
## Scenario 3, C vs. C+R == 0  1.098  0.411  2.67  0.0076 **
## Scenario 4, C vs. C+R == 0  0.359  0.338  1.06  0.2878
## Scenario 5, C vs. C+R == 0  0.231  0.406  0.57  0.5702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

## D.6 By-Change Stimulus Effects

### D.6.1 Stimulus Detection Effects

This appendix details the model-fitting process for explaining probability of a participant detecting a change in terms of change properties and interface experimental condition (Section 5.5.4.1).

```
MELogit.Q1Det1 <- glmer(Hit ~ (1 | Participant) + (1 | TrueScenario)
, data = QLevelLogit, family = binomial, nAGQ = 1)
summary(MELogit.Q1Det1, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: Hit ~ (1 | Participant) + (1 | TrueScenario)
## Data: QLevelLogit
```

```

##
##      AIC      BIC  logLik deviance df.resid
##  1032.3  1046.6  -513.2  1026.3    855
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.1513 -1.0436  0.5166  0.7267  1.3890
##
## Random effects:
##  Groups      Name      Variance Std.Dev.
##  Participant (Intercept) 0.1671  0.4088
##  TrueScenario (Intercept) 1.0601  1.0296
## Number of obs: 858, groups: Participant, 33; TrueScenario, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.1772    0.3568    3.3 0.000968 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

First, a model was fit containing all interactions between measured experimental condition and change / ordering properties. Participant and TrueScenario are included as random effects. Not all effects are significant and the model does not converge.

```

MELogit.Q1Det20 <- glmer(Hit ~ Interface * FirstBooklet
                        + Interface * Cause
                        + Interface * SizeLarge
                        + Interface * Evidence
                        + Interface * Leading
                        + (1 | Participant) + (1 | TrueScenario) # Consider
playing around with random components after checking for huge interactions?
                        , data = QLevellogit, family = binomial, nAGQ = 1)
## Warning in checkConv(attr("opt", "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.00322805
## (tol = 0.001, component 12)
summary(MELogit.Q1Det20, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: Hit ~ Interface * FirstBooklet + Interface * Cause + Interface *
## SizeLarge + Interface * Evidence + Interface * Leading +
## (1 | Participant) + (1 | TrueScenario)
## Data: QLevellogit
##
##      AIC      BIC  logLik deviance df.resid
##   842.3   926.4  -403.1   806.3    774
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.6084 -0.6272  0.2850  0.5529  2.4685
##
## Random effects:
##  Groups      Name      Variance Std.Dev.

```

```

## Participant (Intercept) 0.3612 0.6010
## TrueScenario (Intercept) 0.5338 0.7306
## Number of obs: 792, groups: Participant, 33; TrueScenario, 10
##
## Fixed effects:
##
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.06452 0.53138 -2.003 0.0451 *
## InterfaceC+R -0.36699 0.59698 -0.615 0.5387
## FirstBookletTRUE 0.06608 0.32946 0.201 0.8410
## CauseB -1.03316 0.35379 -2.920 0.0035 **
## CauseC -0.38429 0.42285 -0.909 0.3634
## CauseD -0.80999 0.43925 -1.844 0.0652 .
## SizeLargeTRUE 1.81971 0.33306 5.464 4.66e-08 ***
## Evidence 0.17241 0.03614 4.771 1.83e-06 ***
## LeadingTRUE 0.56082 0.30874 1.816 0.0693 .
## InterfaceC+R:FirstBookletTRUE -0.23614 0.55042 -0.429 0.6679
## InterfaceC+R:CauseB 0.62648 0.46569 1.345 0.1785
## InterfaceC+R:CauseC 0.70951 0.51939 1.366 0.1719
## InterfaceC+R:CauseD 0.34254 0.55376 0.619 0.5362
## InterfaceC+R:SizeLargeTRUE -0.34545 0.40099 -0.861 0.3890
## InterfaceC+R:Evidence 0.02258 0.04514 0.500 0.6170
## InterfaceC+R:LeadingTRUE -0.22648 0.40859 -0.554 0.5794
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

After several stepwise removals (see maximum likelihood comparison below), model 13 contains all significant parameters:

```

MELogit.Q1Det13 <- glmer(Hit ~ Cause
                        + SizeLarge
                        + Evidence
                        + (1 | Participant) + (1 | TrueScenario)
                        , data = QLevelLogit, family = binomial, nAGQ = 1)
summary(MELogit.Q1Det13, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: Hit ~ Cause + SizeLarge + Evidence + (1 | Participant) + (1 |
## TrueScenario)
## Data: QLevelLogit
##
## AIC      BIC    logLik deviance df.resid
## 831.2    868.6  -407.6   815.2    784
##
## Scaled residuals:
## Min      1Q  Median      3Q      Max
## -5.5112 -0.6403  0.2982  0.5632  2.7164
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## Participant (Intercept) 0.3549  0.5957
## TrueScenario (Intercept) 0.6499  0.8062
## Number of obs: 792, groups: Participant, 33; TrueScenario, 10
##

```

```

## Fixed effects:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.91367    0.41113  -2.222  0.0263 *
## CauseB      -0.64508    0.25945  -2.486  0.0129 *
## CauseC       0.02300    0.34130   0.067  0.9463
## CauseD      -0.46875    0.32999  -1.420  0.1555
## SizeLargeTRUE 1.61290    0.26575   6.069 1.29e-09 ***
## Evidence     0.15953    0.02727   5.851 4.89e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##           (Intr) CauseB CauseC CauseD SLTRUE
## CauseB      -0.145
## CauseC      -0.066  0.340
## CauseD       0.004  0.293  0.584
## SizeLrgTRUE -0.216  0.196 -0.328 -0.269
## Evidence    -0.554 -0.192 -0.388 -0.458  0.273

```

Another, simpler model omits Change Cause, leaving only Size and Evidence as significantly predicting the probability of a change being detected:

```

MELogit.Q1Det9 <- glmer(Hit ~ SizeLarge + Evidence
                        + (1 | Participant) + (1|TrueScenario)
                        , data = QLevellogit, family = binomial, nAGQ = 1)
summary(MELogit.Q1Det9, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula:
## Hit ~ SizeLarge + Evidence + (1 | Participant) + (1 | TrueScenario)
## Data: QLevellogit
##
##           AIC          BIC    logLik deviance df.resid
##      834.4      857.8    -412.2    824.4      787
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.6164 -0.6582  0.2773  0.6102  2.3141
##
## Random effects:
## Groups      Name          Variance Std.Dev.
## Participant (Intercept) 0.3407   0.5837
## TrueScenario (Intercept) 0.8755   0.9357
## Number of obs: 792, groups: Participant, 33; TrueScenario, 10
##
## Fixed effects:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.00960    0.44425  -2.273  0.0231 *
## SizeLargeTRUE 1.82970    0.23566   7.764 8.23e-15 ***
## Evidence     0.14493    0.02484   5.834 5.42e-09 ***
## ---

```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
##           (Intr) SLTRUE
## SizeLrgTRUE -0.203
## Evidence    -0.642  0.185
```

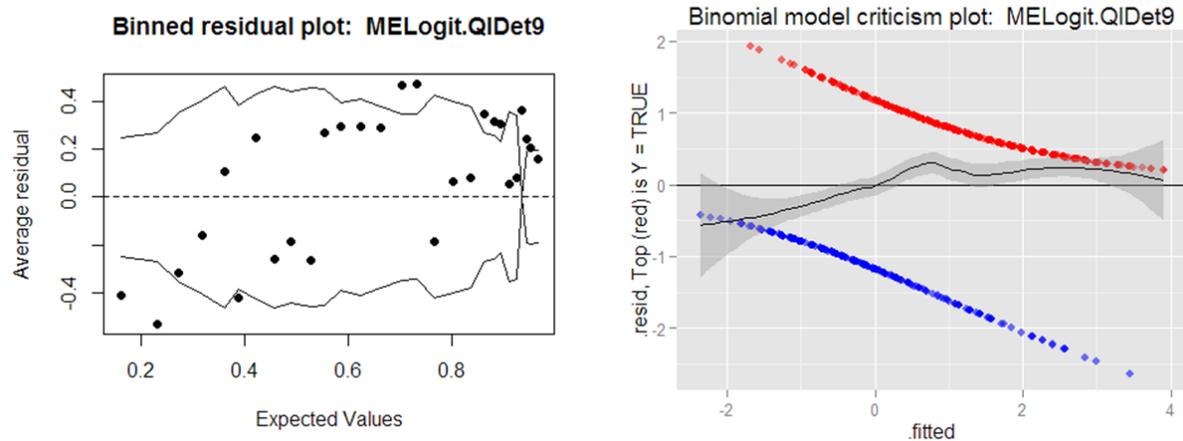
Maximum likelihood comparison suggests that model 13 is a likely fit, and that model 9 is a simpler alternative:

```
print(anova(MELogit.Q1Det1,MELogit.Q1Det9,MELogit.Q1Det10,MELogit.Q1Det13,MELogit.Q1Det14,
MELogit.Q1Det15,MELogit.Q1Det16,MELogit.Q1Det17,MELogit.Q1Det18,MELogit.Q1Det19,
MELogit.Q1Det20))
## Data: QLevelLogit
## Models:
## MELogit.Q1Det1: Hit ~ (1 | Participant) + (1 | TrueScenario)
## MELogit.Q1Det9: Hit ~ SizeLarge + Evidence + (1 | Participant) + (1 |
TrueScenario)
## MELogit.Q1Det10: Hit ~ SizeLarge + Evidence + Leading + (1 | Participant) + (1 |
TrueScenario)
## MELogit.Q1Det10:      TrueScenario)
## MELogit.Q1Det13: Hit ~ Cause + SizeLarge + Evidence + (1 | Participant) + (1 |
TrueScenario)
## MELogit.Q1Det13:      TrueScenario)
## MELogit.Q1Det14: Hit ~ Cause + SizeLarge + Evidence + Leading + (1 | Participant)
+
## MELogit.Q1Det14:      (1 | TrueScenario)
## MELogit.Q1Det15: Hit ~ Interface + Cause + SizeLarge + Evidence + Leading + (1 |
## MELogit.Q1Det15:      Participant) + (1 | TrueScenario)
## MELogit.Q1Det16: Hit ~ FirstBooklet + Interface + Cause + SizeLarge + Evidence +
## MELogit.Q1Det16:      Leading + (1 | Participant) + (1 | TrueScenario)
## MELogit.Q1Det17: Hit ~ FirstBooklet + Interface * Cause + SizeLarge + Evidence +
## MELogit.Q1Det17:      Leading + (1 | Participant) + (1 | TrueScenario)
## MELogit.Q1Det18: Hit ~ FirstBooklet + Interface * Cause + Interface * SizeLarge +
## MELogit.Q1Det18:      Evidence + Leading + (1 | Participant) + (1 | TrueScenario)
## MELogit.Q1Det19: Hit ~ Interface * FirstBooklet + Interface * Cause + Interface *
## MELogit.Q1Det19:      SizeLarge + Evidence + Interface * Leading + (1 |
Participant) +
## MELogit.Q1Det19:      (1 | TrueScenario)
## MELogit.Q1Det20: Hit ~ Interface * FirstBooklet + Interface * Cause + Interface *
## MELogit.Q1Det20:      SizeLarge + Interface * Evidence + Interface * Leading +
## MELogit.Q1Det20:      (1 | Participant) + (1 | TrueScenario)
##
##           Df      AIC      BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## MELogit.Q1Det1  3 1032.32 1046.58 -513.16  1026.32
## MELogit.Q1Det9  5  834.41  857.79 -412.21   824.41 201.9040    2 < 2e-16 ***
## MELogit.Q1Det10 6  835.21  863.25 -411.60   823.21  1.2064    1 0.27204
## MELogit.Q1Det13 8  831.24  868.64 -407.62   815.24  7.9628    2 0.01866 *
## MELogit.Q1Det14 9  829.74  871.82 -405.87   811.74  3.4995    1 0.06139 .
## MELogit.Q1Det15 10 831.60  878.34 -405.80   811.60  0.1455    1 0.70289
## MELogit.Q1Det16 11 833.51  884.93 -405.75   811.51  0.0900    1 0.76412
## MELogit.Q1Det17 14 836.25  901.70 -404.13   808.25  3.2566    3 0.35372
## MELogit.Q1Det18 15 837.13  907.24 -403.56   807.13  1.1255    1 0.28874
## MELogit.Q1Det19 17 840.51  919.98 -403.26   806.51  0.6136    2 0.73578
## MELogit.Q1Det20 18 842.27  926.41 -403.13   806.27  0.2441    1 0.62129
```

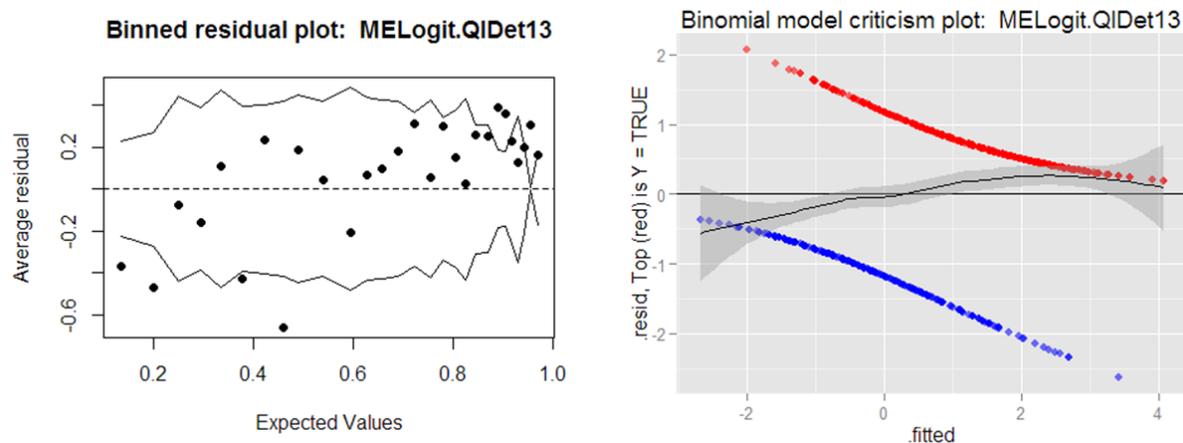
```
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model comparison suggests two candidates: model QlDet9 with the lowest BIC and the more complex QlDet13 with the lowest AIC. Alternative random effect structures were considered, but Participant and True Scenario were the most likely:

```
print(anova(MELogit.QlDet13, MELogit.QlDet13.1, MELogit.QlDet13.2))
## Data: QLevelLogit
## Models:
## MELogit.QlDet13.2: Hit ~ Cause + SizeLarge + Evidence + (1 | Participant)
## MELogit.QlDet13: Hit ~ Cause + SizeLarge + Evidence + (1 | Participant) + (1 |
## MELogit.QlDet13: TrueScenario)
## MELogit.QlDet13.1: Hit ~ Cause + SizeLarge + Evidence + (1 | Order/Participant) +
## MELogit.QlDet13.1: (1 | TrueScenario)
##
##           Df    AIC    BIC logLik deviance Chisq Chi Df    Pr(>Chisq)
## MELogit.QlDet13.2  7 868.28 901.01 -427.14  854.28
## MELogit.QlDet13    8 831.24 868.64 -407.62  815.24 39.041    1 4.151e-10 ***
## MELogit.QlDet13.1  9 833.24 875.31 -407.62  815.24 0.000    1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



**Figure 57 - Residual inspection plot for Model MELogitQIDet9, explaining likelihood of a change being detected. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.**



**Figure 58 - Residual inspection plot for Model MELogitQIDet13, explaining likelihood of a change being detected. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.**

Odds ratios for both simpler and more complex candidate models:

```
print(round(exp(melr.estimatedCI(MELogit.QIDet9)),2))
```

```
##           Est  LL  UL
## (Intercept) 0.36 0.15 0.87
## SizeLargeTRUE 6.23 3.93 9.89
## Evidence     1.16 1.10 1.21
```

```
print(round(exp(melr.estimatedCI(MELogit.QIDet13)),2))
```

```
##           Est  LL  UL
## (Intercept) 0.40 0.18 0.90
## CauseB      0.52 0.32 0.87
```

```
## CauseC      1.02 0.52 2.00
## CauseD      0.63 0.33 1.19
## SizeLargeTRUE 5.02 2.98 8.45
## Evidence    1.17 1.11 1.24
```

The Evidence present in the changes ranged from 2.4 to 22.5 accumulated ‘average days consumption’ units. The effect of half this range on odds of detection, according to the simpler model, is:

```
print(confint(glht(MELogit.Q1Det9, MELogit.Q1Det9.c2)))

## Simultaneous Confidence Intervals

## Fit: glmer(formula = Detected ~ SizeLarge + Evidence + (1 | Participant) +
## (1 | TrueScenario), data = QLevelLogit, family = binomial,
## nAGQ = 1)
##
## Quantile = 1.96
## 95% family-wise confidence level
##
## Linear Hypotheses:
##              Estimate lwr   upr
## Evidence 2.4 vs. 12.45 == 0 1.4565  0.9672 1.9459
##
print(round(exp(coef(glht(MELogit.Q1Det9, MELogit.Q1Det9.c2))),2))
## Evidence 2.4 vs. 12.45
##              4.29
```

## D.6.2 Stimulus Diagnosis Effects

The following logistic regression models describe probability of a change being Rightly Diagnosed (pRD) given that a participant detected it and attempted diagnosis. The model does not need to be formulated with an offset, since the RightDiag variable is N/A for un-detected or un-diagnosed cases.

```
MELogit.Q1Diag1 <- glmer(RightDiag ~ (1|Participant) # offset(DiagGuess)
                        , data = QLevelLogit, family = binomial, nAGQ = 1)
summary(MELogit.Q1Diag1, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: RightDiag ~ (1 | Participant)
## Data: QLevelLogit
##
##      AIC      BIC   logLik deviance df.resid
## 608.2    616.4  -302.1   604.2     465
##
```

```

## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.7322 -0.7322 -0.7322  1.3657  1.3657
##
## Random effects:
## Groups      Name                Variance Std.Dev.
## Participant (Intercept) 4.039e-14 2.01e-07
## Number of obs: 467, groups: Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.62328    0.09708   -6.42 1.36e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

MELogit.Q1Diag19 <- glmer(RightDiag ~ Interface * FirstBooklet
                          + Interface * Leading
                          + Interface * Cause
                          + Interface * SizeLarge
                          + Interface * Evidence
                          + (1|Participant)
                          , data = QLevelLogit, family = binomial, nAGQ = 1)
summary(MELogit.Q1Diag19, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: RightDiag ~ Interface * FirstBooklet + Interface * Leading +
##          Interface * Cause + Interface * SizeLarge + Interface * Evidence +
##          (1 | Participant)
## Data: QLevelLogit
##
##      AIC      BIC    logLik deviance df.resid
##  529.6    598.3   -247.8   495.6     403
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.3909 -0.7242 -0.4356  0.9053  3.0343
##
## Random effects:
## Groups      Name                Variance Std.Dev.
## Participant (Intercept) 0          0
## Number of obs: 420, groups: Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.670342   0.564102  -2.961 0.003066 **
## InterfaceC+R  -0.005861    0.723948  -0.008 0.993541
## FirstBookletTRUE  0.254791    0.337805  0.754 0.450696
## LeadingTRUE     0.279928    0.372086  0.752 0.451858
## CauseB         0.400040    0.600764  0.666 0.505485
## CauseC         2.096398    0.539810  3.884 0.000103 ***
## CauseD         1.866669    0.574173  3.251 0.001150 **
## SizeLargeTRUE  -0.099770    0.388134  -0.257 0.797139

```

```

## Evidence -0.055239 0.032518 -1.699 0.089376 .
## InterfaceC+R:FirstBookletTRUE -0.285931 0.443424 -0.645 0.519041
## InterfaceC+R:LeadingTRUE -0.522145 0.533087 -0.979 0.327346
## InterfaceC+R:CauseB 0.612175 0.782767 0.782 0.434176
## InterfaceC+R:CauseC -1.144226 0.709685 -1.612 0.106896
## InterfaceC+R:CauseD -0.237246 0.773485 -0.307 0.759054
## InterfaceC+R:SizeLargeTRUE -0.197505 0.533948 -0.370 0.711461
## InterfaceC+R:Evidence 0.108383 0.043743 2.478 0.013223 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

After stepwise removal, these two models are candidates:

```

MELogit.QlDiag13 <- glmer(RightDiag ~ Interface * Cause
                        + Interface * Evidence
                        + (1|Participant)
                        , data = QLevellogit, family = binomial, nAGQ = 1)
summary(MELogit.QlDiag13, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula:
## RightDiag ~ Interface * Cause + Interface * Evidence + (1 | Participant)
## Data: QLevellogit
##
##      AIC      BIC   logLik deviance df.resid
##  519.8   564.2  -248.9   497.8     409
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.4153 -0.7450 -0.4414  0.8801  2.6602
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## Participant (Intercept) 4e-14    2e-07
## Number of obs: 420, groups: Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.51396   0.42912  -3.528 0.000419 ***
## InterfaceC+R -0.26072   0.59252  -0.440 0.659919
## CauseB        0.31910   0.59235   0.539 0.590095
## CauseC        2.04265   0.52918   3.860 0.000113 ***
## CauseD        1.76593   0.55880   3.160 0.001576 **
## Evidence      -0.04451   0.03011  -1.478 0.139314
## InterfaceC+R:CauseB 0.54133   0.74982   0.722 0.470328
## InterfaceC+R:CauseC -1.19811   0.68801  -1.741 0.081611 .
## InterfaceC+R:CauseD -0.29150   0.73791  -0.395 0.692822
## InterfaceC+R:Evidence 0.09718   0.04050   2.400 0.016408 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:

```

```
##          (Intr) IntC+R CauseB CauseC CauseD Evidnc IC+R:CB IC+R:CC IC+R:CD
## InterfacC+R -0.724
## CauseB      -0.528  0.382
## CauseC      -0.439  0.318  0.510
## CauseD      -0.378  0.274  0.492  0.719
## Evidence     -0.456  0.330 -0.101 -0.447 -0.506
## IntrfC+R:CB  0.417 -0.558 -0.790 -0.403 -0.388  0.080
## IntrfC+R:CC  0.337 -0.449 -0.393 -0.769 -0.553  0.344  0.535
## IntrfC+R:CD  0.286 -0.365 -0.372 -0.545 -0.757  0.383  0.504  0.702
## IntrfC+R:E   0.339 -0.502  0.075  0.332  0.376 -0.743 -0.064 -0.391 -0.468
```

Model MELogit.QlDiag13 has interaction terms moderately significant  $p < .082$ . However, evidence for the Cause \* Interface interactions does not meet the  $p < .05$  convention. Further stepwise removal and likelihood ratio tests offer a simpler model with main effects all significant,  $p < .009$ :

```
MELogit.QlDiag10 <- glmer(RightDiag ~ Cause
                          + Interface
                          + (1|Participant)
                          , data = QLevelLogit, family = binomial, nAGQ = 1)
summary(MELogit.QlDiag10, corr = FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: RightDiag ~ Cause + Interface + (1 | Participant)
## Data: QLevelLogit
##
##      AIC      BIC   logLik deviance df.resid
##  566.7    591.6   -277.3   554.7     461
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.1784 -0.7813 -0.5280  0.9750  2.4866
##
## Random effects:
## Groups      Name                Variance Std.Dev.
## Participant (Intercept) 0                0
## Number of obs: 467, groups: Participant, 33
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.8218    0.2422  -7.522 5.39e-14 ***
## CauseB       0.6133    0.3259   1.882 0.0598 .
## CauseC       1.3282    0.2771   4.794 1.64e-06 ***
## CauseD       1.6058    0.2873   5.590 2.28e-08 ***
## InterfaceC+R 0.5443    0.2078   2.620 0.0088 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of fixed effects could have been required in summary()
##
## Correlation of Fixed Effects:
```

```
##          (Intr) CauseB CauseC CauseD
## CauseB   -0.547
## CauseC   -0.652  0.487
## CauseD   -0.650  0.469  0.552
## InterfacC+R -0.501 -0.019 -0.005  0.036
```

Note that in Model 10, Participant does not account for any additional variation as a random variable.

```
print(anova(MELogit.QlDiag1, MELogit.QlDiag10, MELogit.QlDiag11, MELogit.QlDiag12,
MELogit.QlDiag13,
MELogit.QlDiag14,MELogit.QlDiag15,MELogit.QlDiag16,MELogit.QlDiag17,MELogit.QlDiag18,
MELogit.QlDiag19))
## Data: QLevelLogit
## Models:
## MELogit.QlDiag1: RightDiag ~ (1 | Participant)
## MELogit.QlDiag10: RightDiag ~ Cause + Interface + (1 | Participant)
## MELogit.QlDiag11: RightDiag ~ Cause + Interface + Evidence + (1 | Participant)
## MELogit.QlDiag12: RightDiag ~ Cause + Interface * Evidence + (1 | Participant)
## MELogit.QlDiag13: RightDiag ~ Interface * Cause + Interface * Evidence + (1 |
Participant)
## MELogit.QlDiag14: RightDiag ~ Leading + Interface * Cause + Interface * Evidence +
## MELogit.QlDiag14:      (1 | Participant)
## MELogit.QlDiag15: RightDiag ~ Interface * Leading + Interface * Cause + Interface
*
## MELogit.QlDiag15:      Evidence + (1 | Participant)
## MELogit.QlDiag16: RightDiag ~ Interface * Leading + Interface * Cause + Interface
*
## MELogit.QlDiag16:      Evidence + SizeLarge + (1 | Participant)
## MELogit.QlDiag17: RightDiag ~ FirstBooklet + Interface * Leading + Interface *
## MELogit.QlDiag17:      Cause + Interface * Evidence + SizeLarge + (1 | Participant)
## MELogit.QlDiag18: RightDiag ~ Interface * FirstBooklet + Interface * Leading +
## MELogit.QlDiag18:      Interface * Cause + Interface * Evidence + SizeLarge + (1 |
## MELogit.QlDiag18:      Participant)
## MELogit.QlDiag19: RightDiag ~ Interface * FirstBooklet + Interface * Leading +
## MELogit.QlDiag19:      Interface * Cause + Interface * SizeLarge + Interface *
Evidence +
## MELogit.QlDiag19:      (1 | Participant)
##          Df      AIC      BIC  loglik deviance   Chisq Chi Df      Pr(>Chisq)
## MELogit.QlDiag1   2 608.16 616.45 -302.08   604.16
## MELogit.QlDiag10  6 566.67 591.55 -277.34   554.67 49.4838      4 4.628e-10 ***
## MELogit.QlDiag11  7 523.16 551.44 -254.58   509.16 45.5126      1 1.517e-11 ***
## MELogit.QlDiag12  8 521.43 553.75 -252.72   505.43  3.7287      1 0.05349 .
## MELogit.QlDiag13 11 519.76 564.20 -248.88   497.76  7.6689      3 0.05337 .
## MELogit.QlDiag14 12 521.62 570.10 -248.81   497.62  0.1448      1 0.70354
## MELogit.QlDiag15 13 522.92 575.44 -248.46   496.92  0.7013      1 0.40236
## MELogit.QlDiag16 14 524.35 580.91 -248.17   496.35  0.5686      1 0.45080
## MELogit.QlDiag17 15 526.17 586.78 -248.09   496.17  0.1754      1 0.67533
## MELogit.QlDiag18 16 527.75 592.39 -247.87   495.75  0.4232      1 0.51534
## MELogit.QlDiag19 17 529.61 598.30 -247.81   495.61  0.1372      1 0.71106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 13 has the lowest AIC, but depends on an Interface \* Cause interaction  $p > .08$ . Model 10 explains diagnosis likelihood more simply in terms of Interface and Change cause.

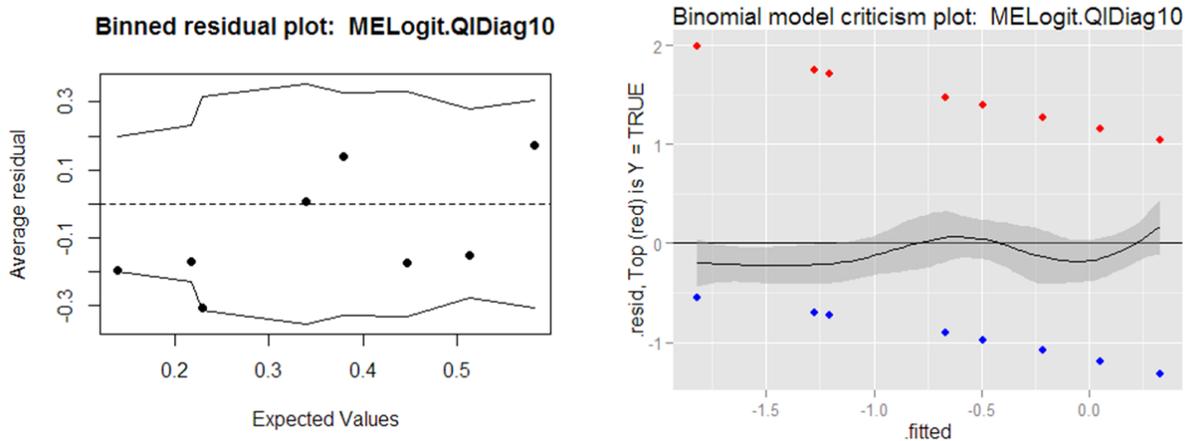
The simpler model has more significant (and simpler to interpret) terms:

```
print(round(exp(melr.estimatedCI(MELogit.QlDiag10)),2))
##           Est   LL   UL
## (Intercept) 0.16 0.10 0.26
## CauseB      1.85 0.97 3.50
## CauseC      3.77 2.19 6.50
## CauseD      4.98 2.84 8.75
## InterfaceC+R 1.72 1.15 2.59
```

Some alternate random effects were tested by maximum likelihood comparison, and showed no improvement:

```
anova(MELogit.QlDiag10, MELogit.QlDiag10.1, MELogit.QlDiag10.2)
## Data: QLevelLogit
## Models:
## MELogit.QlDiag10: RightDiag ~ Cause + Interface + (1 | Participant)
## MELogit.QlDiag10.1: RightDiag ~ Cause + Interface + (1 | TrueScenario)
## MELogit.QlDiag10.2: RightDiag ~ Cause + Interface + (1 | Participant) + (1 |
TrueScenario)
##           Df    AIC    BIC  logLik deviance Chisq Chi Df    Pr(>Chisq)
## MELogit.QlDiag10    6 566.67 591.55 -277.34  554.67
## MELogit.QlDiag10.1  6 566.67 591.55 -277.34  554.67    0    0 <2e-16 ***
## MELogit.QlDiag10.2  7 568.67 597.70 -277.34  554.67    0    1 1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The simpler model seems to have reasonably distributed residuals.



**Figure 59- Residual inspection plot for Model MELogit.QIDdiag10, explaining likelihood of a participant rightly diagnosing a change that they detected and attempted a diagnosis for. Binned residuals (left) and LOESS smoothed average residual with 95% confidence intervals (right) shown.**

## D.7 Outside Influences / Assumption checking

### D.7.1 Data Entry Validation

A random 10% sample of the marked changes was selected, and paper charts (e.g. Figure 29) re-scored by an industrial engineering graduate student according to the interpretation rules outlined in Section 5.2.7. The original and independent ratings were compared for inter-rater reliability using Krippendorff's Alpha (Krippendorff, 2004) and are summarized below in Table 30. The worst 20 mismatches were re-inspected to determine whether experimenter or independent validator had made an error transcribing. Of the twenty, three were ambiguous and seventeen were errors by the independent validator. This suggests the experimental data was accurately transcribed.

**Table 30 - Inter-Rater Reliability test of Experimental II data coding. 127 samples coded by 2 raters.**

	Krippendorff's test type	$\alpha$	Re-inspection
<i>Marked Date (text)</i>	Nominal	.967	One ambiguous, one validator error
<i>Marked Confidence</i>	Ordinal	.989	Three validator errors
<i>Marked Direction</i>	Nominal	.969	One ambiguous
<i>Marked Cause</i>	Nominal	.975	Two validator errors
<i>Box Location</i>	Nominal	.988	One ambiguous
<i>Line Start (mm)</i>	Ratio	.986	Three validator errors
<i>X Mark (mm)</i>	Ratio	.952	Three validator errors
<i>Line End (mm)</i>	Ratio	.986	Three validator errors

### D.7.2 Comparing Scoring Rules -Main effects

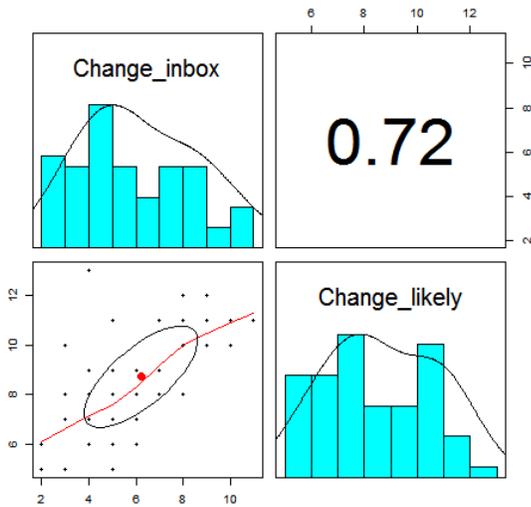
Hits and Right Diagnoses for both “In Box” and “Likely” scoring rules (Section 5.2.8) were calculated, and compared in Section 5.5.5.2.

```
temp <- data.frame(XError =
as.vector(na.omit(c(QLevelFactors$XError_likely,QLevelFactors$XError_inbox)))
, Rule = as.factor(c(rep.int("Likely",
length(na.omit(QLevelFactors$XError_likely))),rep.int("InBox",
length(na.omit(QLevelFactors$XError_inbox))))))

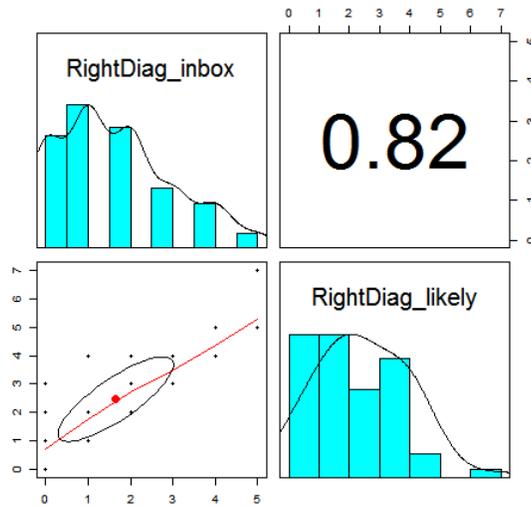
describeBy(temp, temp$Rule)

## group: InBox
##      vars   n mean   sd median trimmed  mad  min max range skew
## XError   1 410 -9.06 32.9   -2   -5.34 10.38 -238 131  369 -2.1
## Rule*    2 410  1.00  0.0    1    1.00  0.00   1   1    0  NaN
## -----
## group: Likely
##      vars   n mean   sd median trimmed  mad  min max range skew
## XError   1 576 30.5 87.08    2   17.96 22.24 -238 555  793  2.04
## Rule*    2 576  2.0  0.00    2    2.00  0.00   2   2    0  NaN
```

Comparing MT&amp;R Change Detection Scoring Rules, by Participant



Comparing MT&amp;R Change Diagnosis Scoring Rules, by Participant



**Figure 60 - Pearson Correlations and scatter plots comparing "InBox" and "Likely" scoring rules for Detection (left) and Diagnosis (right). Scores are aggregated by-participant.**

```
print(RuleReliability.Change) # Looks reasonable!

##
## Reliability analysis Comparing M&T Change Detection Scoring Rules
## Call: alpha(x = PLevelFactors[, ImportantVars], title = "Comparing M&T Change
## Detection Scoring Rules")
##
## raw_alpha std.alpha G6(smc) average_r S/N ase mean sd
## 0.83 0.84 0.72 0.72 5.1 0.16 7.5 2.1
##
## lower alpha upper 95% confidence boundaries
## 0.51 0.83 1.14
##
## Reliability if an item is dropped:
## raw_alpha std.alpha G6(smc) average_r S/N alpha se
## Change_inbox 0.72 0.72 0.51 0.72 NA NA
## Change_likely 0.72 0.72 0.51 0.72 NA NA
##
## Item statistics
## n r r.cor r.drop mean sd
## Change_inbox 66 0.93 0.78 0.72 6.2 2.4
## Change_likely 66 0.93 0.78 0.72 8.7 2.0
```

```
print(RuleReliability.Diag)

##
## Reliability analysis Comparing Change Diagnosis Scoring Rules
```

```
## Call: alpha(x = PLevelFactors[, ImportantVars], title = "Comparing Change
Diagnosis Scoring Rules")
##
##   raw_alpha std.alpha G6(smc) average_r S/N ase mean sd
##     0.9      0.9      0.82      0.82 8.8 0.15 2.1 1.4
##
## lower alpha upper      95% confidence boundaries
## 0.61 0.9 1.18
##
## Reliability if an item is dropped:
##           raw_alpha std.alpha G6(smc) average_r S/N alpha se
## RightDiag_inbox      0.82      0.82      0.67      0.82 NA      NA
## RightDiag_likely      0.82      0.82      0.67      0.82 NA      NA
##
## Item statistics
##           n      r r.cor r.drop mean sd
## RightDiag_inbox 66 0.95 0.86 0.82 1.7 1.4
## RightDiag_likely 66 0.95 0.86 0.82 2.5 1.5
##
## Non missing response frequency for each item
##           0 1 2 3 4 5 7 miss
## RightDiag_inbox 0.23 0.29 0.24 0.12 0.09 0.03 0.00 0
## RightDiag_likely 0.09 0.18 0.27 0.17 0.23 0.05 0.02 0
```

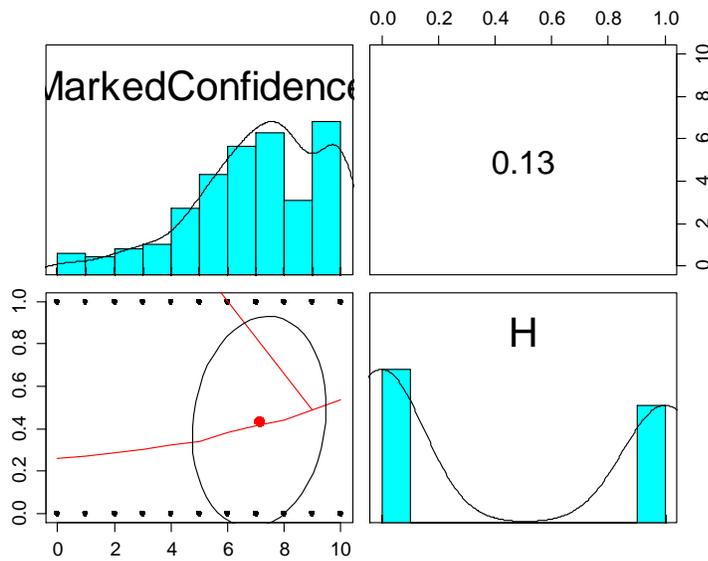
Interface condition made no difference in percentage of responses scored as hits.

```
temp <- subset(QLevelFactors, select=c(Change_inbox, Change_likely, Interface,
Participant))
temp <- mutate(temp, HInbox = !is.na(Change_inbox), HLikely = !is.na(Change_likely))

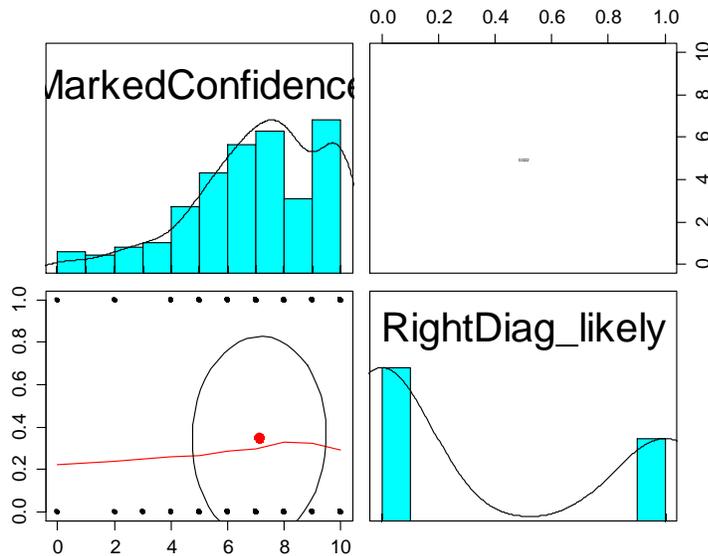
describeBy(temp, temp$Interface)
## group: C
##           vars  n mean  sd median trimmed  mad min max range
## Interface*    3 667 1.00 0.00      1   1.00 0.00  1  1  0
## HInbox*       5 667 0.31 0.46      0   0.26 0.00  0  1  1
## HLikely*      6 667 0.43 0.50      0   0.42 0.00  0  1  1
## -----
## group: C+R
##           vars  n mean  sd median trimmed  mad min max range
## Interface*    3 662 2.00 0.00      2   2.00 0.00  2  2  0
## HInbox*       5 662 0.31 0.46      0   0.26 0.00  0  1  1
## HLikely*      6 662 0.43 0.50      0   0.42 0.00  0  1  1
```

### D.7.3 Confidence Correlations

The Marked Confidence of a response seemed to have little association with correctness. As shown below, confidence was weakly correlated with likelihood of the response hitting a true change, and uncorrelated with the likelihood of a hit being rightly diagnosed.



**Figure 61 - By-Response (n=1317) correlation between indicated confidence in the change being new, and whether response Hit a true change.**



**Figure 62 - By-Response (n=463) correlation between indicated confidence in the change being new, and whether Hits with Diagnosis attempts were Rightly Diagnosed.**

#### D.7.4 Comparing rates of Attempted Diagnosis

The propensity of participants to diagnose each type of change cause in attempts was examined using un-aggregated data of each response.

#### D.7.4.1 Proportion of Hits and Diagnosis Attempts by Interface condition

For responses that “hit” any of 12 true changes (omitting the extra-large change 2A / 7A), there was no evidence of a difference between interface conditions in the proportions of types of changes “hit”:

```

misdiagD.table <- CrossTable(misdiag$TrueCause, misdiag$Interface, digits = 2,
prop.t=FALSE, prop.r=FALSE, chisq=TRUE, expected=TRUE) #, format="SPSS")
##
##
##   Cell Contents
##   -----|
##           N
##   Expected N
##   Chi-square contribution
##           N / Col Total
##   -----|
##
##
## Total Observations in Table:  515
##
##
##   misdiag$TrueCause | misdiag$Interface
##   misdiag$TrueCause |      C      |      C+R      | Row Total |
##   -----|-----|-----|-----|
##           A         |      71      |      61      |     132   |
##           |      66.38   |      65.62   |           |
##           |      0.32    |      0.32    |           |
##           |      0.27    |      0.24    |           |
##   -----|-----|-----|-----|
##           B         |      48      |      51      |     99    |
##           |      49.79   |      49.21   |           |
##           |      0.06    |      0.06    |           |
##           |      0.19    |      0.20    |           |
##   -----|-----|-----|-----|
##           C         |      73      |      77      |    150   |
##           |      75.44   |      74.56   |           |
##           |      0.08    |      0.08    |           |
##           |      0.28    |      0.30    |           |
##   -----|-----|-----|-----|
##           D         |      67      |      67      |    134   |
##           |      67.39   |      66.61   |           |
##           |      0.00    |      0.00    |           |
##           |      0.26    |      0.26    |           |
##   -----|-----|-----|-----|
##           Column Total |      259     |      256     |    515   |
##           |      0.50    |      0.50    |           |
##   -----|-----|-----|-----|
##
##
## Statistics for All Table Factors
##
##

```

```
## Pearson's Chi-squared test
## -----
## Chi^2 = 0.9377076    d.f. = 3    p = 0.8163199
##
```

However, considering the propensity of participants to attempt each possible diagnosis (including false alarms), there were significant differences in proportions, both between interface conditions and from the proportion of true causes present. Omitting changes matched to extra-large changes 2A / 7A, there were equal proportion of the four true cause types present.

```
misdiagHM.table <- CrossTable(misdiag$MarkedCause, misdiag$Interface, digits = 2,
prop.t=FALSE, prop.r=FALSE, chisq=TRUE, expected=TRUE) #, format="SPSS")
```

```
##
##
##      Cell Contents
## |-----|
## |                N |
## |      Expected N |
## | Chi-square contribution |
## |      N / Col Total |
## |-----|
##
##
## Total Observations in Table:  1024
##
##
##      misdiag$MarkedCause | misdiag$Interface
##      -----|-----|-----|-----|
##      A |      68 |      95 |      163 |
##      |  77.52 |  85.48 |          |
##      |   1.17 |   1.06 |          |
##      |   0.14 |   0.18 |          |
##      -----|-----|-----|
##      B |      68 |     112 |      180 |
##      |  85.61 |  94.39 |          |
##      |   3.62 |   3.28 |          |
##      |   0.14 |   0.21 |          |
##      -----|-----|-----|
##      C |     208 |     146 |      354 |
##      | 168.36 | 185.64 |          |
##      |   9.33 |   8.47 |          |
##      |   0.43 |   0.27 |          |
##      -----|-----|-----|
##      D |     143 |     184 |      327 |
##      | 155.52 | 171.48 |          |
##      |   1.01 |   0.91 |          |
##      |   0.29 |   0.34 |          |
##      -----|-----|-----|
##      Column Total |     487 |     537 |      1024 |
##      |   0.48 |   0.52 |          |
##      -----|-----|-----|
```

```
##
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## -----
## Chi^2 = 28.85477    d.f. = 3    p = 2.40244e-06
##
```

Both experimental conditions induced diagnosis attempts in proportions significantly different than expected by the proportion of true causes present in each scenario set (three of change types A..D).

```
print(chisq.test(misdiagHM.table$t[, "C"]
                , p = rep(0.25,4)))
##
## Chi-squared test for given probabilities
##
## data:  misdiagHM.table$t[, "C"]
## X-squared = 112.269, df = 3, p-value < 2.2e-16

print(chisq.test(misdiagHM.table$t[, "C+R"]
                , p = rep(0.25,4)))
##
## Chi-squared test for given probabilities
##
## data:  misdiagHM.table$t[, "C+R"]
## X-squared = 34.6276, df = 3, p-value = 1.46e-07
```

#### D.7.4.2 Cross-Tabulating Misdiagnoses

Ignoring the 1/5 of responses without a diagnosis and the 1/2 of responses scored as False Alarms, the cross-tabulation of attempts to true causes (omitting change 2A / 7A) is shown below:

```
misdiag.table <- CrossTable(misdiag$MarkedCause, misdiag$TrueCause, digits = 2,
prop.t=FALSE, prop.chisq = FALSE) # , format="SPSS")
##
##
##      Cell Contents
## |-----|
## |                N |
## |      N / Row Total |
## |      N / Col Total |
## |-----|
##
##
## Total Observations in Table: 420
##
```

```
##
##
## misdiag$MarkedCause | misdiag$TrueCause
## -----|-----|-----|-----|-----|
##          A |          B |          C |          D | Row Total |
##          19 |          8 |          20 |          10 |          57 |
##          0.33 |         0.14 |         0.35 |         0.18 |         0.14 |
##          0.17 |         0.10 |         0.16 |         0.09 |          |
## -----|-----|-----|-----|-----|
##          B |          22 |          23 |          16 |          7 |          68 |
##          0.32 |         0.34 |         0.24 |         0.10 |         0.16 |
##          0.20 |         0.29 |         0.13 |         0.07 |          |
## -----|-----|-----|-----|-----|
##          C |          34 |          27 |          57 |          34 |         152 |
##          0.22 |         0.18 |         0.38 |         0.22 |         0.36 |
##          0.31 |         0.34 |         0.46 |         0.32 |          |
## -----|-----|-----|-----|-----|
##          D |          35 |          21 |          32 |          55 |         143 |
##          0.24 |         0.15 |         0.22 |         0.38 |         0.34 |
##          0.32 |         0.27 |         0.26 |         0.52 |          |
## -----|-----|-----|-----|-----|
##          Column Total |          110 |          79 |          125 |          106 |         420 |
##          0.26 |         0.19 |         0.30 |         0.25 |          |
## -----|-----|-----|-----|-----|
##
##
```

For only the C interface condition:

```
misdiagC.table <- CrossTable(misdiagC$MarkedCause, misdiagC$TrueCause, digits = 2,
prop.t=FALSE, prop.chisq = FALSE) #, format="SPSS")
##
##
##      Cell Contents
## -----|-----|
##              N |
##      N / Row Total |
##      N / Col Total |
## -----|-----|
##
##
## Total Observations in Table: 194
##
##
## misdiagC$MarkedCause | misdiagC$TrueCause
## -----|-----|-----|-----|-----|
##          A |          8 |          4 |          9 |          6 |          27 |
##          0.30 |         0.15 |         0.33 |         0.22 |         0.14 |
##          0.14 |         0.11 |         0.17 |         0.12 |          |
## -----|-----|-----|-----|-----|
##          B |          7 |          6 |          7 |          2 |          22 |
##          0.32 |         0.27 |         0.32 |         0.09 |         0.11 |
##          0.12 |         0.17 |         0.13 |         0.04 |          |
## -----|-----|-----|-----|-----|
##          C |          20 |          15 |          25 |          22 |          82 |
##          0.24 |         0.18 |         0.30 |         0.27 |         0.42 |
## -----|-----|-----|-----|-----|
##
```

##		0.35	0.43	0.47	0.45	
##	-----	-----	-----	-----	-----	-----
##	D	22	10	12	19	63
##		0.35	0.16	0.19	0.30	0.32
##		0.39	0.29	0.23	0.39	
##	-----	-----	-----	-----	-----	-----
##	Column Total	57	35	53	49	194
##		0.29	0.18	0.27	0.25	
##	-----	-----	-----	-----	-----	-----
##						

In only the C+R Interface condition:

```

misdiagCR.table <- CrossTable(misdiagCR$MarkedCause, misdiagCR$TrueCause, digits = 2,
prop.t=FALSE, prop.chisq = FALSE) #, format="SPSS")
##
##
## Cell Contents
## |-----|
## |                N |
## |      N / Row Total |
## |      N / Col Total |
## |-----|
##
##
## Total Observations in Table:  226
##
##
##
## misdiagCR$MarkedCause | misdiagCR$TrueCause
##-----|-----|-----|-----|-----|
##      A |      B |      C |      D | Row Total |
##-----|-----|-----|-----|-----|
##      11 |      4 |      11 |      4 |      30 |
##      0.37 | 0.13 | 0.37 | 0.13 | 0.13 |
##      0.21 | 0.09 | 0.15 | 0.07 |      |
##-----|-----|-----|-----|-----|
##      15 |      17 |      9 |      5 |      46 |
##      0.33 | 0.37 | 0.20 | 0.11 | 0.20 |
##      0.28 | 0.39 | 0.12 | 0.09 |      |
##-----|-----|-----|-----|-----|
##      14 |      12 |      32 |      12 |      70 |
##      0.20 | 0.17 | 0.46 | 0.17 | 0.31 |
##      0.26 | 0.27 | 0.44 | 0.21 |      |
##-----|-----|-----|-----|-----|
##      13 |      11 |      20 |      36 |      80 |
##      0.16 | 0.14 | 0.25 | 0.45 | 0.35 |
##      0.25 | 0.25 | 0.28 | 0.63 |      |
##-----|-----|-----|-----|-----|
##      Column Total |      53 |      44 |      72 |      57 |      226 |
##      0.23 | 0.19 | 0.32 | 0.25 |      |
##-----|-----|-----|-----|-----|
##
##

```

## D.7.4.3 Proportion of Diagnosis Attempts between Hits and False Alarms

Comparing the proportion of attempts of each diagnosis cause (A,B,C,D) for hit responses compared to false alarms, the proportions are not significantly different at  $p < .05$ :

```

misdiagFA.table <- CrossTable(misdiag$MarkedCause, is.na(misdiag$TrueCause), digits =
2, prop.t=FALSE, prop.r=FALSE, chisq=TRUE, expected=TRUE) #, format="SPSS")
##
##
##      Cell Contents
## |-----|
## |                N |
## |      Expected N |
## | Chi-square contribution |
## |      N / Col Total |
## |-----|
##
## Total Observations in Table:  1024
##
##
##      misdiag$MarkedCause | is.na(misdiag$TrueCause)
##      FALSE |      TRUE | Row Total |
## -----|-----|-----|
##      A |      57 |      106 |      163 |
##      66.86 |      96.14 |
##      1.45 |      1.01 |
##      0.14 |      0.18 |
## -----|-----|-----|
##      B |      68 |      112 |      180 |
##      73.83 |     106.17 |
##      0.46 |      0.32 |
##      0.16 |      0.19 |
## -----|-----|-----|
##      C |     152 |      202 |      354 |
##     145.20 |     208.80 |
##      0.32 |      0.22 |
##      0.36 |      0.33 |
## -----|-----|-----|
##      D |     143 |      184 |      327 |
##     134.12 |     192.88 |
##      0.59 |      0.41 |
##      0.34 |      0.30 |
## -----|-----|-----|
##      Column Total |     420 |      604 |      1024 |
##      0.41 |      0.59 |
## -----|-----|-----|
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## -----
## Chi^2 =  4.78028      d.f. =  3      p =  0.1886115
##

```

#### D.7.4.4 Correlation between Diagnosis performance and Attempt tendency, by Interface

The relationship between tendency to attempt diagnosis with each change type (HDs) and probability of diagnosing each change type (RDr) is discussed in Section 5.5.5.4. The correlations are significant, though the assumptions of independence for Pearson's product-moment correlation tests are not completely met. While each of 33 participants is independent, their repeated measures in both scenario / interface conditions will be related, and the 4 data points that describe each participant's diagnosis attempts (HDs) for each change type are dependent as they must sum to 100%.

```
with(misdiag.PLevelHrate[which(misdiag.PLevelHrate$Interface == "C"),], cor.test(HDs,
RDr))

##
## Pearson's product-moment correlation
##
## data: ADs and pRD
## t = 12.2288, df = 108, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.6704 0.8307
## sample estimates:
## cor
## 0.7620

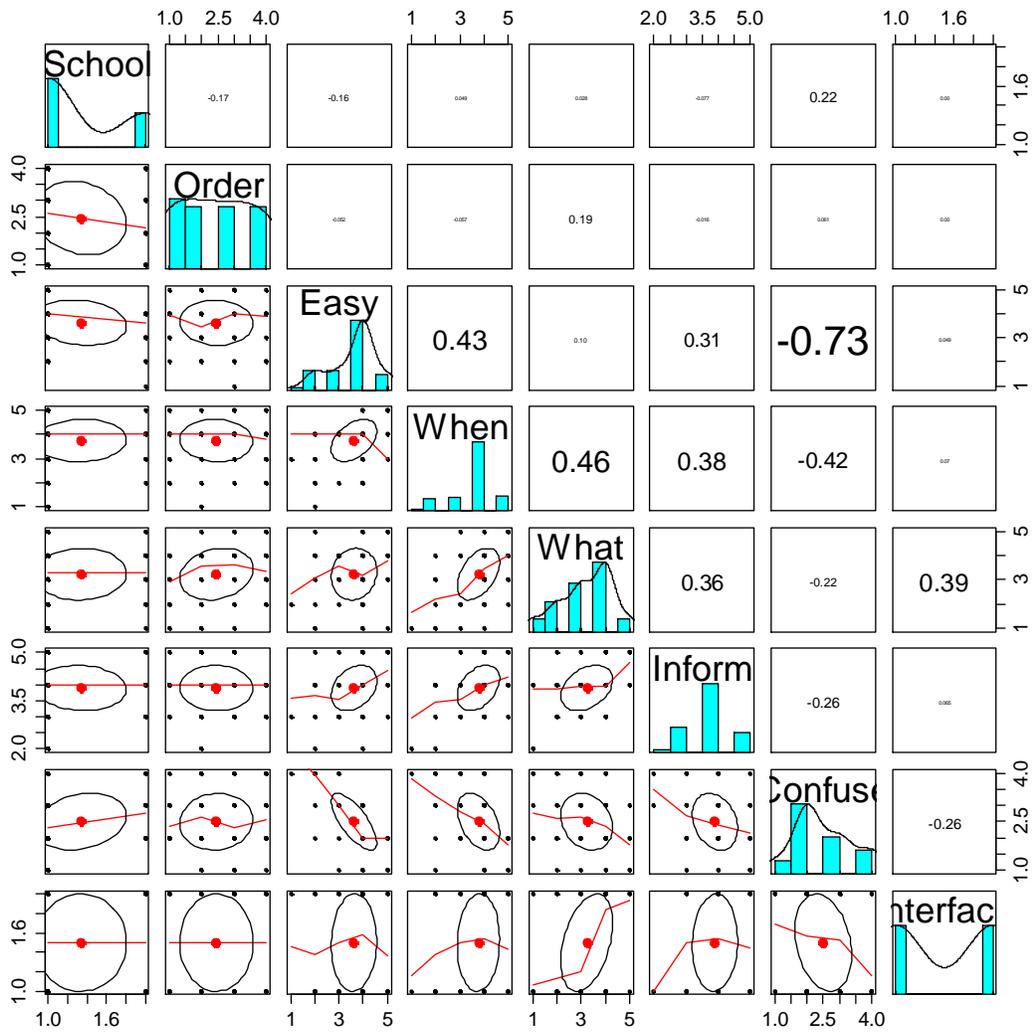
with (misdiag.PLevelHrate[which(misdiag.PLevelHrate$Interface == "C+R"),],
cor.test(HDs, RDr))

##
## Pearson's product-moment correlation
##
## data: ADs and pRD
## t = 9.3944, df = 121, p-value = 4.441e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.5337601 0.7412407
## sample estimates:
## cor
## 0.6494
```

### D.7.5 Participant Feedback

#### D.7.5.1 Questionnaire Correlations

Questionnaire responses are described in Section 5.5.6.1. Correlations between questionnaire elements are shown below in Figure 63.



**Figure 63 - Questionnaire responses for all participants ( $N = 33$ ), on questions whether each Interface was Easy, showed When, showed What, was Informative, & was Confusing. Pearson correlations shown in top right.**

## Appendix E Supplementary Analyses

Secondary results from the M&T chart reading Experiment II are collected here, and referred to in the main text.

### E.1 Experiment II By-Scenario Results

#### E.1.1 Detection Scenario Effects

As expected, Detection performance varied with Scenarios (as they contained different numbers of changes, described in section 5.2.2). Detection performance is summarized in Table 31 and Figure 64.

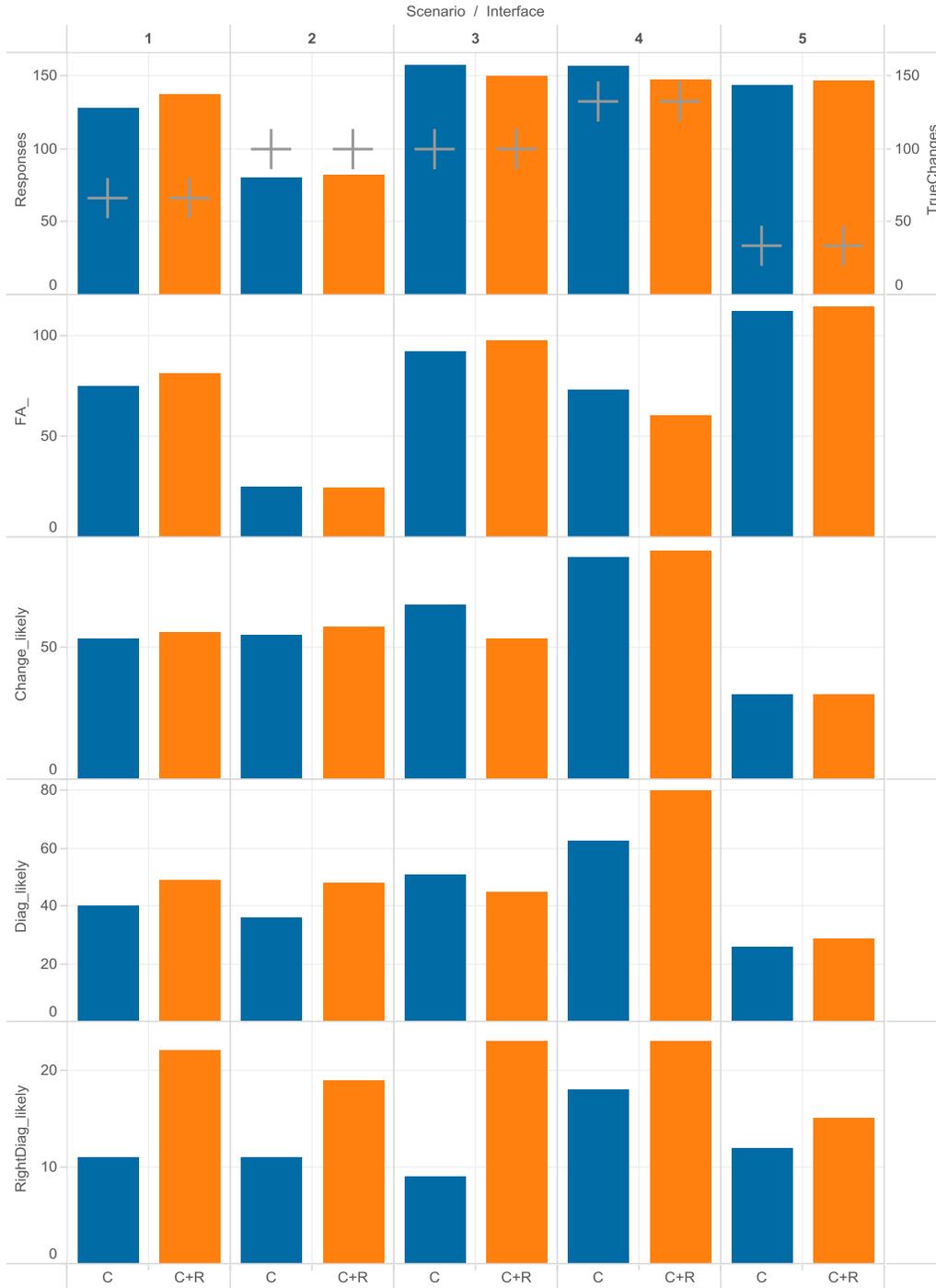
**Table 31 - Detection Hits and False Alarms (“Likely” scoring rule), by Scenario (aggregating G<sub>1..5</sub> and G<sub>5..10</sub> scenario sets). Mean and Standard Deviation summaries shown.**

<i>Scenario</i>	<b>True Changes present</b>	<b>Mean # of Hits (H)</b>	<b>SD</b>	<b>Mean # of False Alarms (FA)</b>	<b>SD</b>	<b>Hit Rate (Hr)</b>	<b>False Alarm Rate (FAr)</b>
<i>1 (&amp; 6)</i>	2	1.65	0.62	2.36	1.47	82%	59%
<i>2 (&amp; 7)</i>	3	1.71	0.82	0.74	0.93	57%	30%
<i>3 (&amp; 8)</i>	3	1.80	0.71	2.87	1.90	60%	61%
<i>4 (&amp; 9)</i>	4	2.59	0.82	2.02	1.40	65%	44%
<i>5 (&amp; 10)</i>	1	0.97	0.17	3.42	1.55	97%	78%

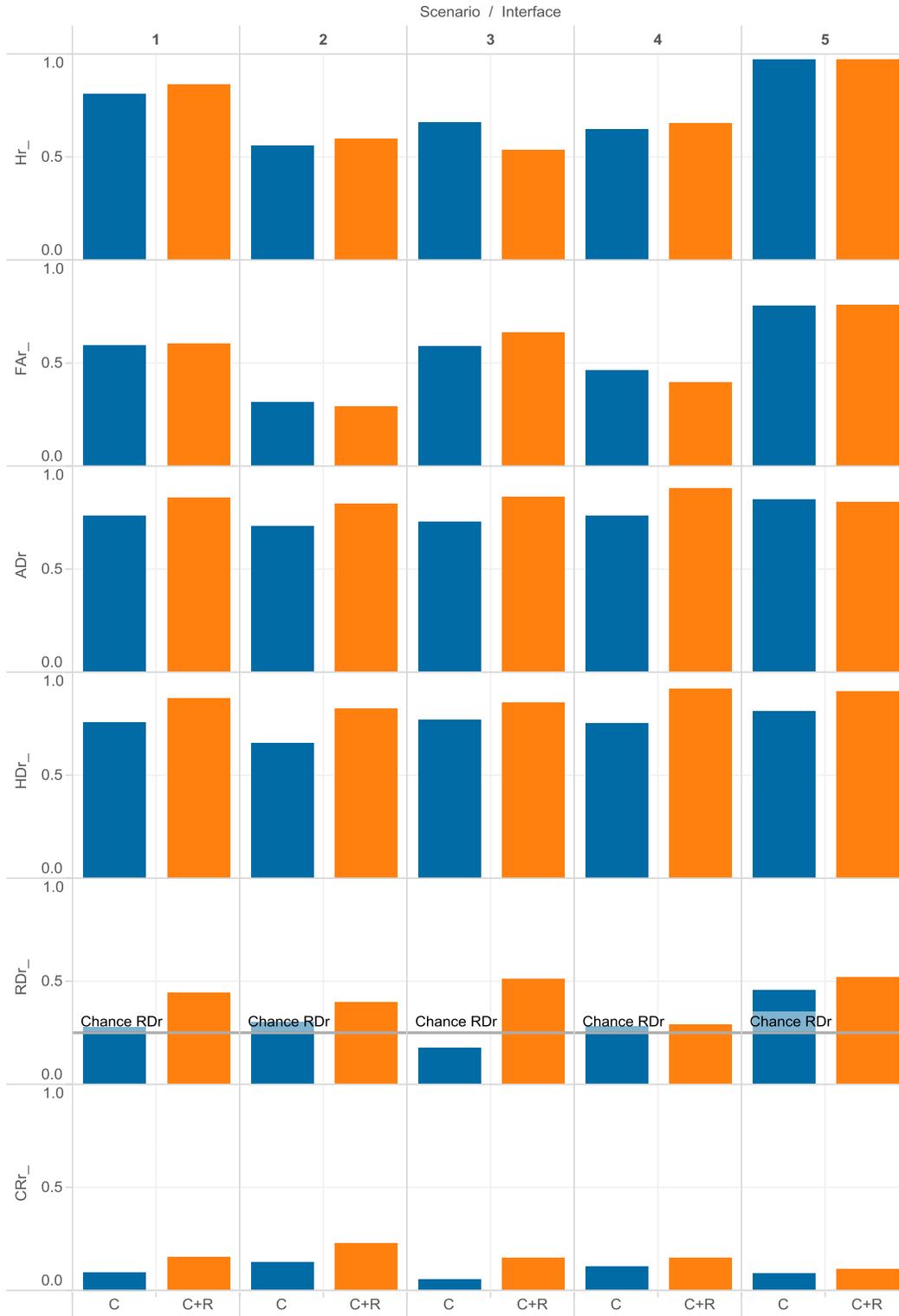
Observations include:

- The number of responses to Scenario 2, which had the large (30%) initial change 2A, was roughly half those of other scenarios.
- Scenarios 3,4, and 5 had comparable response rates despite Scenario 5 containing just one change rather than three or four.
- Hit Rate varied with scenario, with Scenarios 5 and 1 higher than others. This may be expected, due to fewer changes present.

- False Alarm rate (as proportion of responses) varied widely. Scenario 2 was the lowest, Scenario 5 the highest.
- Consistent with interface having no significant effect at participant-level aggregation, (Section 5.5.3.1), by-scenario hit and false alarm rate did not substantially differ between interfaces.



**Figure 64 – Participant total ( $n=33$ ) counts of Responses, False Alarms, Hits, Hits w/ Diagnosis, and Right Diagnoses, by Interface condition. Scenarios combine 5 normal  $G_{1.5}$  and inverted  $G_{6.10}$  TrueScenarios. Number of changes plotted as + symbols: two in scenario 1, two in scenarios 2,3, three in 4, and one in 5.**



**Figure 65 – Average ( $n=33$ ) ratios of Detection and Diagnosis measures, by Interface type and five Scenarios. From top: Hit, False Alarm, Attempted Diagnosis, Hit w/ Diagnosis, Right Diagnosis, and Correct Response ratios described in Table 17.**

### E.1.2 Diagnosis Scenario Effects

Examining scenario type main and interaction effects on diagnosis performance, participants' ~80% propensity to mark a cause and attempt a diagnosis (ADr) was practically equivalent across scenarios (Figure 65). Similarly, the rate of hits containing diagnoses (HDr, Figure 65) was practically equivalent across scenarios. Diagnosis measures that showed practically significant variation between scenarios are listed in Table 32.

**Table 32 – Diagnosis Marked Causes and Right Diagnoses (“Likely” scoring rule), by Scenario (aggregating Ga and Gb scenario sets, n=33). Average of performance in C and C+R interface condition.**

<i>Scenario</i>	<b>True Changes present</b>	<b>Average # of Hits with a Diagnosis Attempt (HD)</b>	<b>S.D.</b>	<b>Average # of Right Diagnoses (RD)</b>	<b>S.D.</b>	<b>Probability of a Hit with Attempted Diagnosis being Right (RDr)</b>	<b>Probability of a Response being a Right Diagnosis (CRr)</b>
1 (& 6)	2	1.35	0.73	0.48	0.56	36%	12%
2 (& 7)	3	1.27	0.83	0.45	0.66	35%	18%
3 (& 8)	3	1.47	0.86	0.52	0.61	35%	11%
4 (& 9)	4	2.17	1.00	0.64	0.76	29%	14%
5 (& 10)	1	0.83	0.38	0.41	0.50	49%	9%

In particular, the probability of a hit with diagnosis being right (RDr) varied from 18% to 52% in both scenario and interface conditions (Figure 65, and Table 32). The most likely mixed-effects binomial logit regression model described RDr probability in terms of fixed effects of Interface  $z = 2.8$ ,  $p = .004$ , and Scenario  $|z| < 1.5$ ,  $p > .15$ , with Participant as a random effect (see Appendix D.5.2). The model indicated that RDr in scenarios 2-5 was not significantly different from scenario 1, both as a main effect or interacting with Interface effects.

Regardless, using a mixed effect binomial logit regression model (see D.5.3) with main and interaction effects of Scenario and Interface found significant differences between particular levels of Interface and Scenario. While scenarios were expected *a priori* to vary in difficulty, p-values should be interpreted with caution, since they were not corrected for familywise error. In the C interface condition, right diagnoses of hits (RDr) were less likely in Scenario 3 (see Figure 65),  $z = -2.03$ ,  $p = .04$  and more likely in Scenario 5,  $z = 2.12$ ,  $p = .03$ , than the average of other

scenarios. Similarly, in the C+R interface condition, RDr was lower in Scenario 4,  $z = -2.66$ ,  $p = .008$ , than average. The model did not find evidence that the C+R interface improved probability of right diagnosis in Scenarios 2, 4 or 5 ( $p > .39$ ).

### E.1.3 Overall Scenario effects

As depicted in Figure 65 and Table 32, probability of a response being completely correct (CRr) varied across scenarios. Mixed-effect logit regression models (see Appendix D.5.4) again found Scenario effects only significant at  $z = 1.71$ ,  $p > .09$ . Given the experimental structure, a model with a main effect of Interface,  $z = 1.81$ ,  $p = .07$  and Scenario,  $|z| < 1.17$ ,  $p > .24$ , with interaction effects,  $|z| < 0.85$ ,  $p > .40$ , and Participant as a random factor was nevertheless used to calculate the same by-scenario contrasts as used to evaluate diagnosis in Section E.1.1.

Even without correcting for familywise error, for the C interface condition, the likelihood of responses being completely correct in each scenario were not distinguishable from average,  $|z| < 1.73$ ,  $p > .08$ . In the C+R condition, participants may have performed better than average in Scenario 2,  $z = 2.10$ ,  $p = .036$ , and worse than average in Scenario 5,  $z = -2.04$ ,  $p = .042$ . Overall the effect of the C+R interface on correct response rate was most apparent in Scenario 3,  $z = 2.67$ ,  $p = .008$ , consistent with the diagnosis-only analysis in Section E.1.1.

## Appendix F Experimental Materials

Experiment II materials are attached below:

### F.1 Experiment II Scenario Definitions

Experiment II scenarios are described in Section 5.2.2. The five scenarios are reiterated below in Table 33, with the exact in-experiment dates. The first 365 days of data were used for training the energy performance regression model. Changes were introduced subsequently, as early as day 425 (60 days into the scenario) for change 2B. Uninterrupted duration of influence (time between changes), and accumulated evidence (estimated effect of the change magnitude over the uninterrupted duration of influence) are tabulated as well.

**Table 33 - Scenario definitions for Experiment II, by number and description (left). Changes (at right) are identified by scenario and number (e.g. 1A, 1B), and defined by their onset time (day of dataset), magnitude of parameter change, uninterrupted duration of influence, and accumulated evidence.**

Scenario	Desc		Change A	Change B	Change C	Change D	Change E
1	"Easy to Mistake"	<i>Parameter</i>	GenCoef	Base			
		<i>Time</i>	545	1050			
		<i>Magnitude</i>	10%	-10%			
		<i>Uninterrupted</i>	505	412			
		<i>Accumulated</i>	15.02	-9.95			
2	"Hidden Diagonal"	<i>Parameter</i>	Base	HDDCoef	GenCoef		
		<i>Time</i>	425	895	1300		
		<i>Magnitude</i>	-50%	10%	-10%		
		<i>Uninterrupted</i>	470	405	162		
		<i>Accumulated</i>	-56.72	10.02	-5.45		
3	"Overlapping Changes"	<i>Parameter</i>	Base	GenCoef	WorkdayCoef		
		<i>Time</i>	700	750	1050		
		<i>Magnitude</i>	20%	-20%	10%		
		<i>Uninterrupted</i>	50	300	412		
		<i>Accumulated</i>	2.41	-18.89	9.93		
4	"Lots of Changes"	<i>Parameter</i>	WorkdayCoef	Base	HDDCoef	WorkdayCoef	
		<i>Time</i>	410	610	1005	1255	
		<i>Magnitude</i>	-10%	10%	-20%	20%	
		<i>Uninterrupted</i>	200	395	250	207	
		<i>Accumulated</i>	-4.91	9.54	-4.88	9.90	
5	"Stereotypical"	<i>Parameter</i>	HDDCoef				
		<i>Time</i>	571				
		<i>Magnitude</i>	10%				
		<i>Uninterrupted</i>	891				
		<i>Accumulated</i>	22.51				

## F.2 Experiment II Instructions

Assume you're an energy manager responsible for the energy performance of 10 office buildings.

Each uses natural gas to:

- Operate a natural gas-fueled backup electric generator from time to time
- Heat the offices in winter. The buildings have a Heating Degree Day (HDD) balance point of 12°C
- Heat hot water and steam for equipment and employees on workdays

The sites have been monitored 4 years, with the first (July 2009 – 2010) used as a modeling baseline.

In each scenario, you'll view charts of a different site's energy performance over the last 3 years. Many things have happened at each site during this time, and your job is to help site staff learn which have affected energy conservation goals.

Your job is to find

- How each site's energy performance has changed since the baseline year
- When energy performance change events most likely happened.
- What might have caused each change event.

You can suggest a diagnosis to help site staff confirm (or repair) the cause of each change, if you're sure.

Remember, a single change event can have long-lasting effects on energy performance! If a steam boiler goes out of tune, for example, it will waste energy all winter and the performance effects might reappear each winter until it is repaired.

Try to judge when changes first happened, and mark each change once when you first notice its effects.

## Instructions

1. Take as much time as you need.
2. Don't guess if you're not sure – behave like a professional energy manager.
3. Don't worry about “right” answers. We want to learn how useful the charts are. Your responses will help us improve energy management software tools.
4. Please use pencil and pen to complete the response sheet for each scenario:
  - a. The pencil for guidelines, scratch marks, or comments.
  - b. The **red pen** to mark your **final responses** on the chart and sheet.
5. For each Scenario, mark energy performance changes on the charts at the date you suspect they happened. Mark as many changes as you judge important.
  - a. Draw **one “X”** to mark **each change**, on the chart you think most clearly shows the change.
  - b. Add a flat line through each “X” long enough to cover the date range when the change **could** have happened.
  - c. Label each “X” mark with a letter (e.g. “A” for the first, “B” second etc.)
6. For each energy performance change, write the letter you used to mark the change (e.g. “A”) on the response sheet, then fill in:
  - a. How confident you are that this chart change shows a **new** performance change, on a scale from zero to ten, where:
    - i. 0 = You shouldn't mark this as a new change
    - ii. 5 = Coin flip odds that this is a new change
    - iii. 10 = Completely confident that it's a new change
  - b. Whether the change, relative to the baseline year:
    - i. (-) = **saved** energy, a good decrease and performance improvement
    - ii. (+) = **increased** energy consumption, a bad overconsumption
  - c. A cause diagnosis that best explains the energy performance change.
    - i. Mark “Not Certain / Can't Tell” if you're not sure - don't guess!
7. When you've marked all the changes in each scenario, please double-check that your answers are clearly marked in red pen. Ask the assistant if you're not sure.
8. The activity has 10 Scenarios, divided into two groups of 5. When you have finished with the first 5, please ask for the second booklet.





## F.4 Sample Experiment II Questionnaire

Thank you for completing the study!

Please mark whether you agree or disagree with these statements:

	Strongly Disagree	Disagree	Un-decided	Agree	Strongly Agree
CUSUM charts were easy to understand					
CUSUM charts clearly showed <b>when</b> changes happened					
CUSUM charts clearly showed <b>what type</b> of changes happened					
CUSUM charts were informative					
CUSUM charts were confusing					
Parameter (RE) charts were easy to understand					
Parameter (RE) charts clearly showed <b>when</b> changes happened					
Parameter (RE) charts clearly showed <b>what type</b> of changes happened					
Parameter (RE) charts were informative					
Parameter (RE) charts were confusing					

Any other comments about this experiment?

## F.5 Experimental Booklets

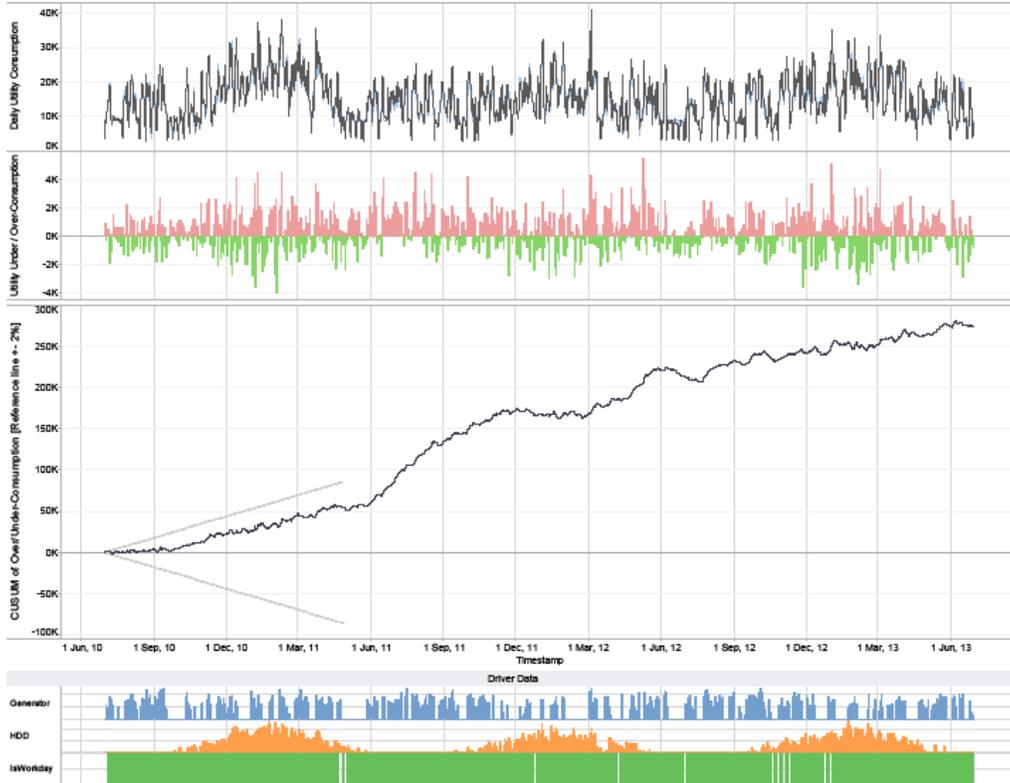
Experimental booklets were printed on 11 x 17" paper. Four varieties of experimental booklet were developed, one for each counterbalanced Scenario x Interface condition (see Section 5.2.3).

In the interests of space, only the booklets for experimental group "A" is presented below. The first booklet A combines the CUSUM-only treatment condition with the set of five scenarios ( $T_C$ ,  $G_{1..5}$ ). The second booklet A presents the CUSUM+RE treatment condition with the inverted set of the same five scenarios ( $T_{C+R}$ ,  $G_{6..10}$ ).

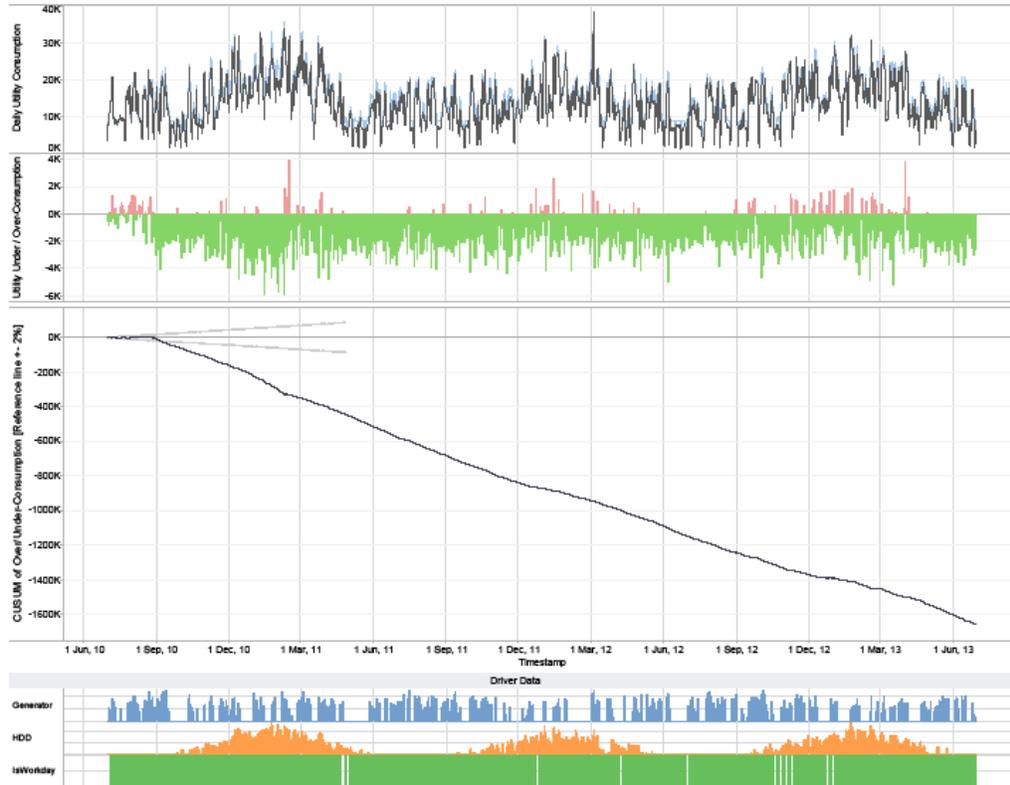
The experimental booklets presented, in order from top to bottom:

- 1) actual and modeled synthetic energy consumption
- 2) Control Chart difference between the two
- 3) CUSUM chart integral
- 4) Driver charts
- 5) To 8) Recursive Estimates charts for the three drivers and intercept (baseload).

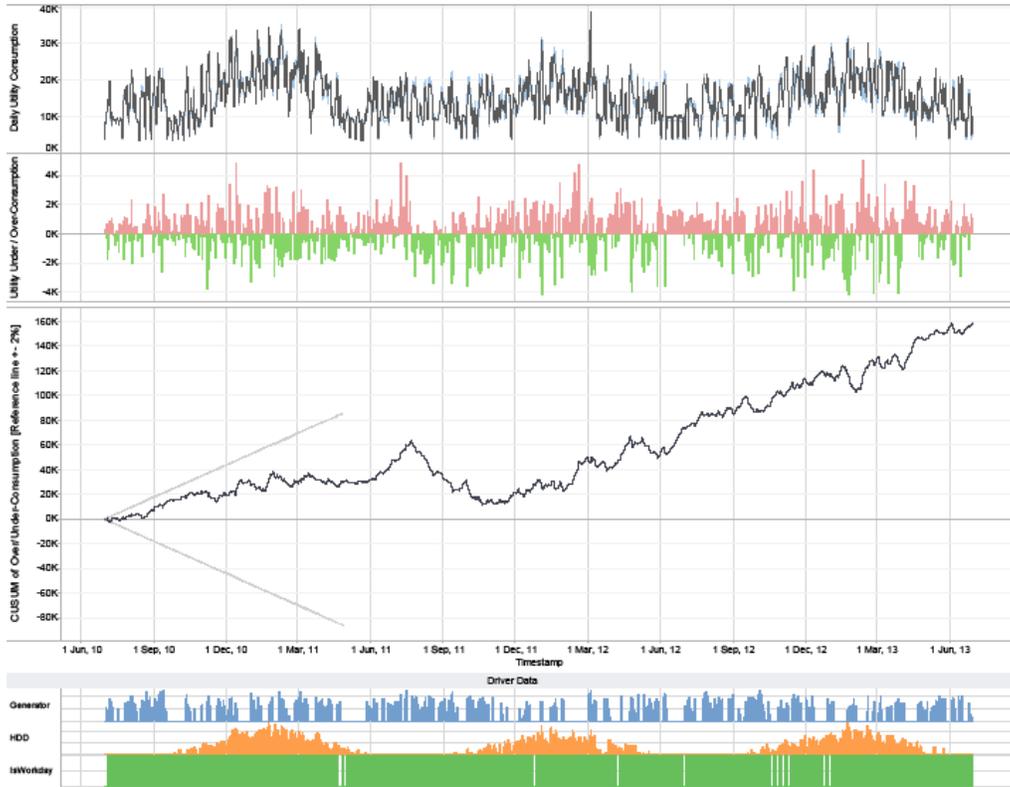
Scenario # 1 of 10



### Scenario # 2 of 10



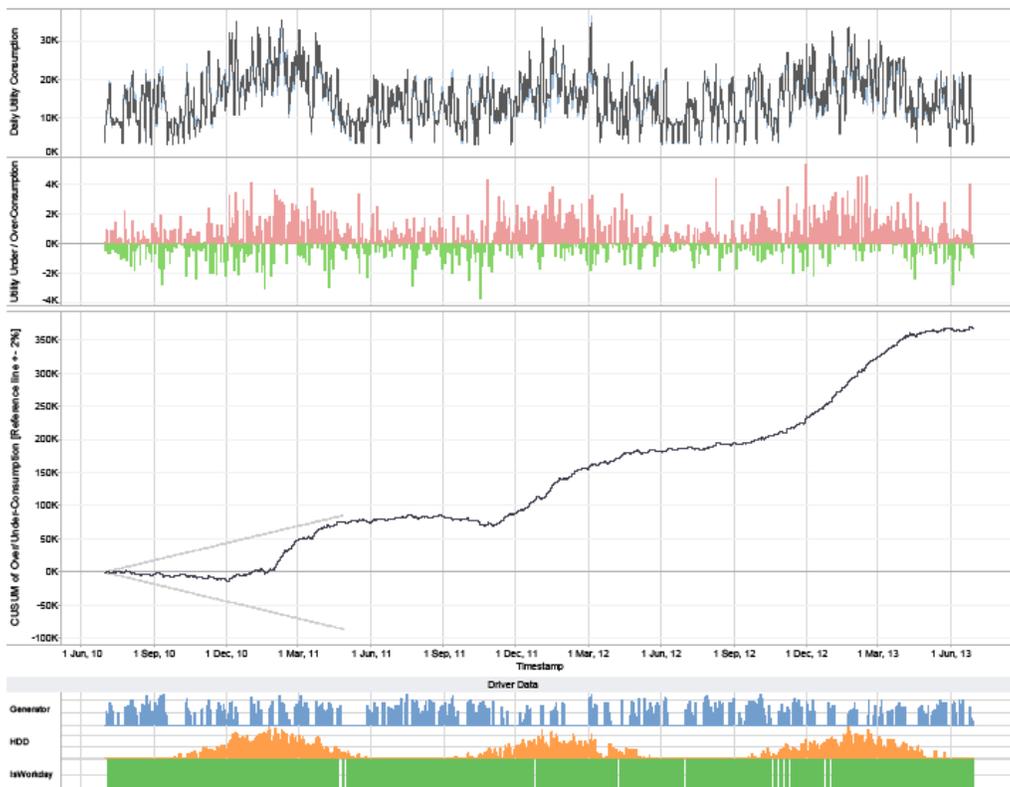
Scenario # 3 of 10



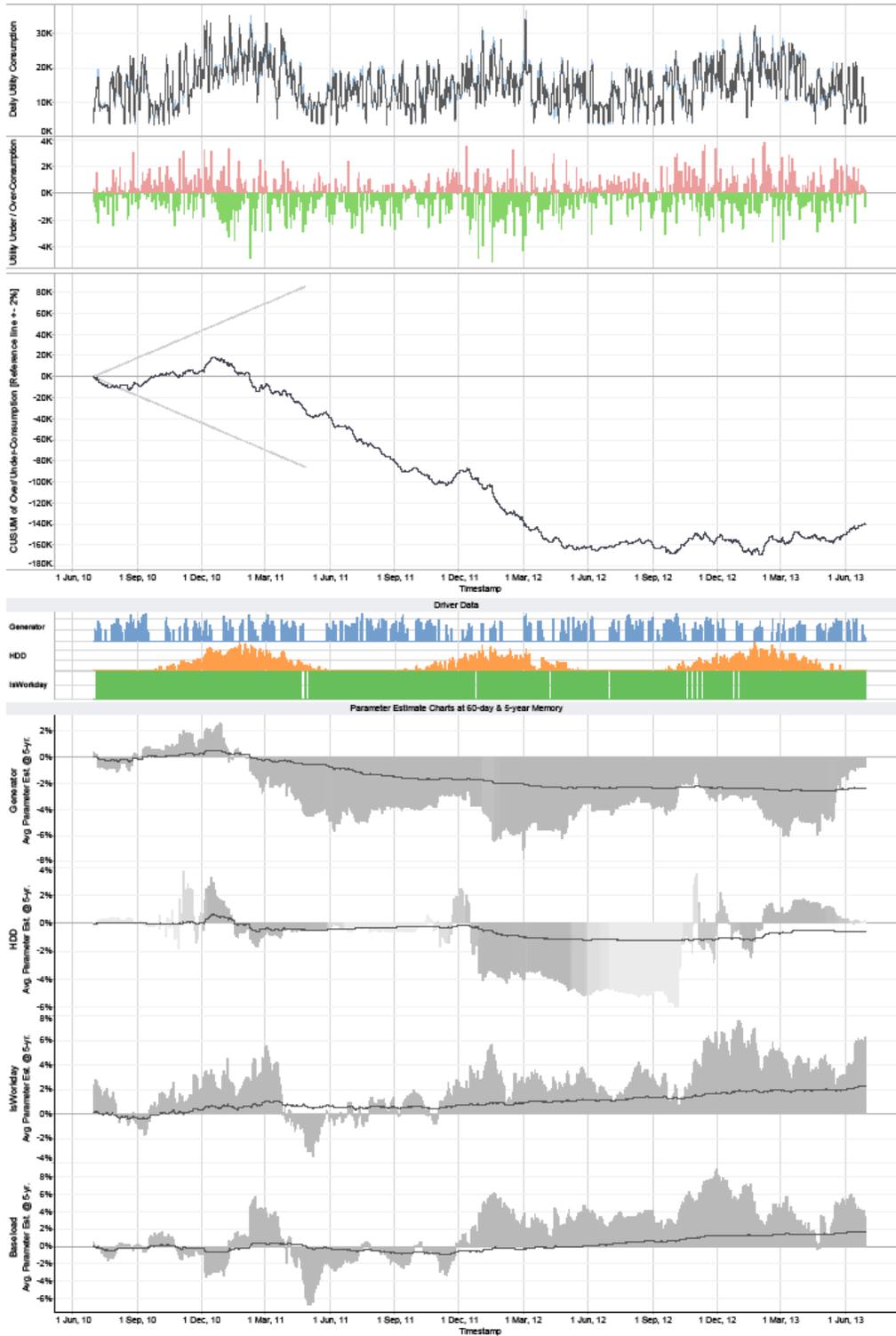
Scenario # 4 of 10



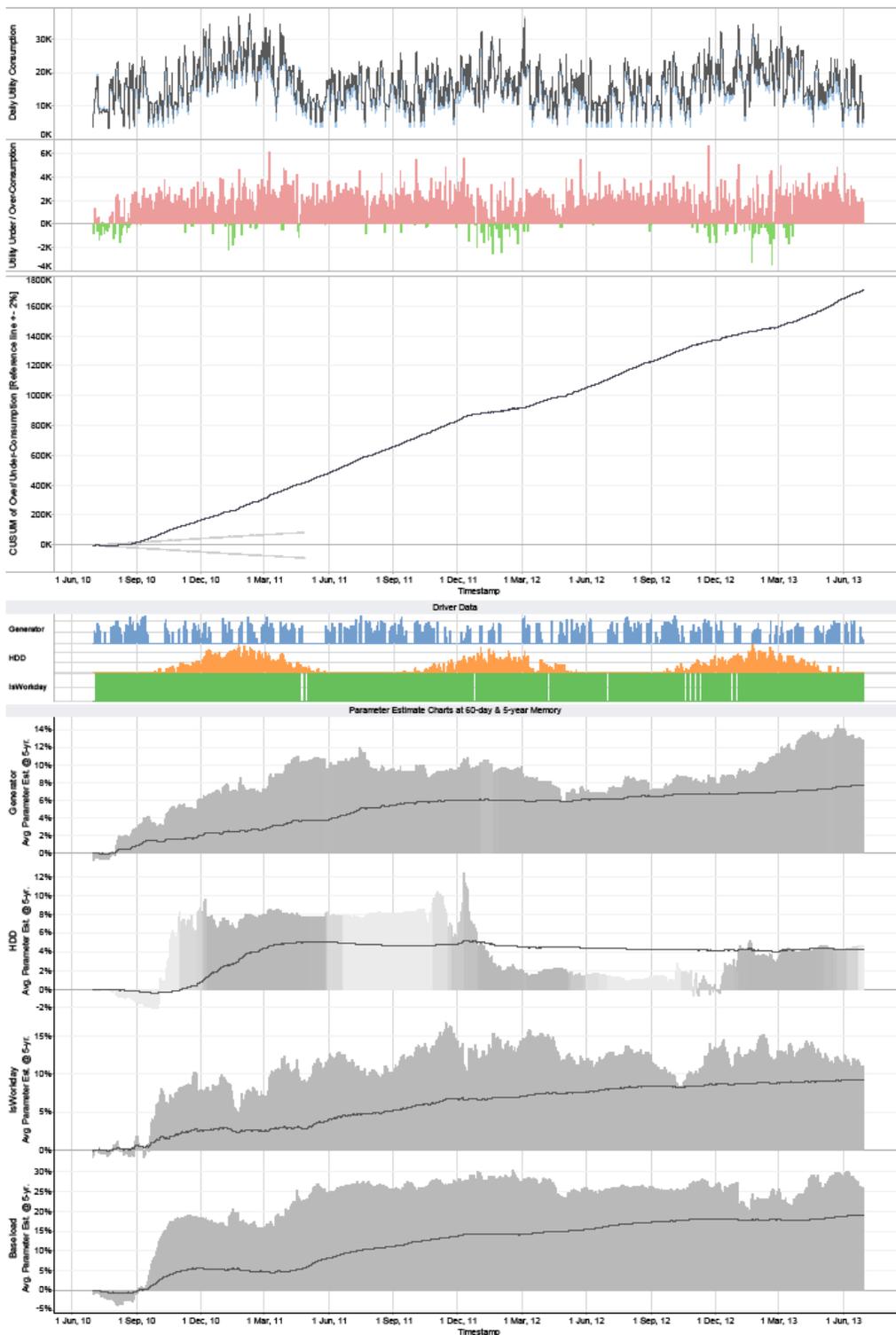
### Scenario # 5 of 10



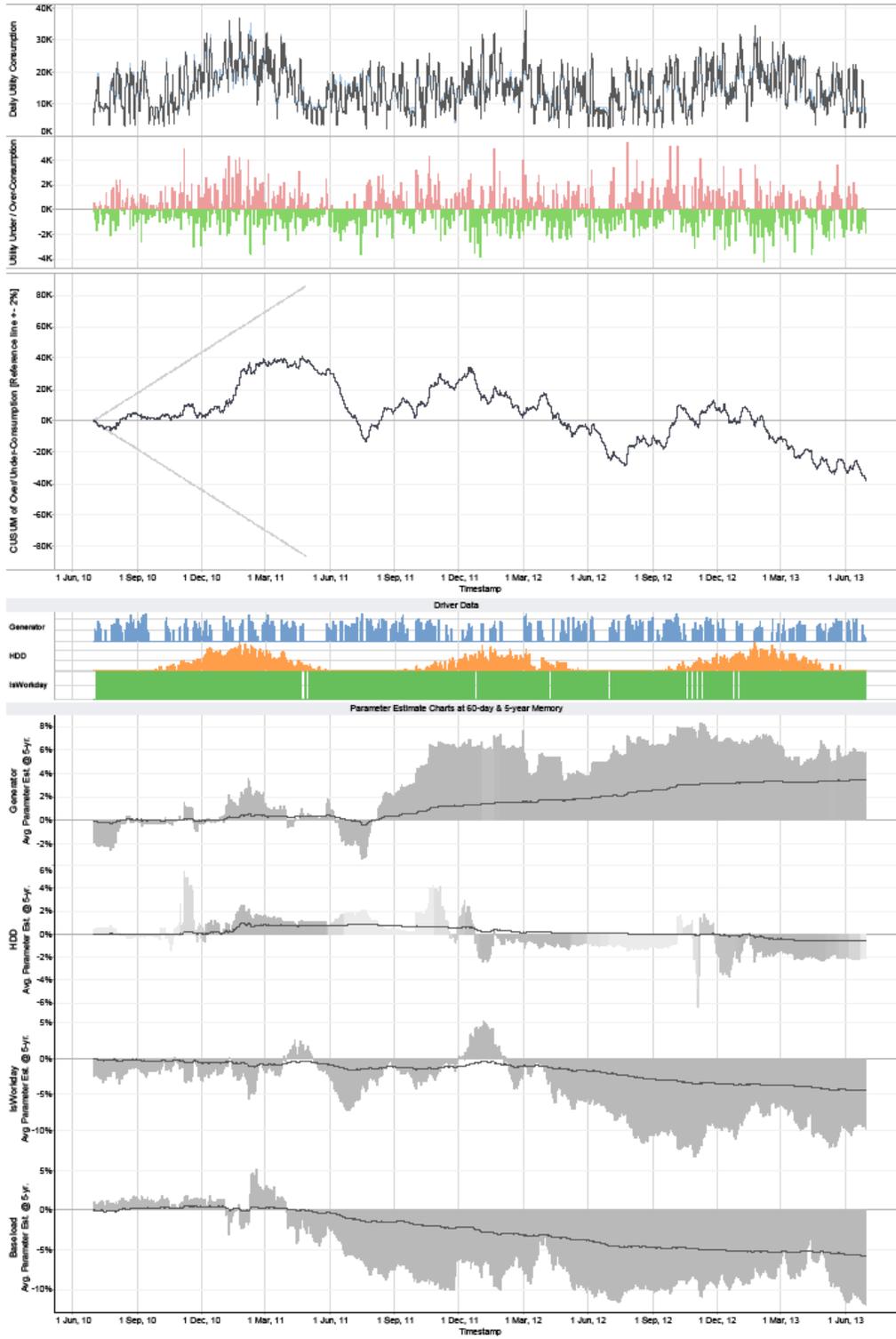
Scenario # 6 of 10



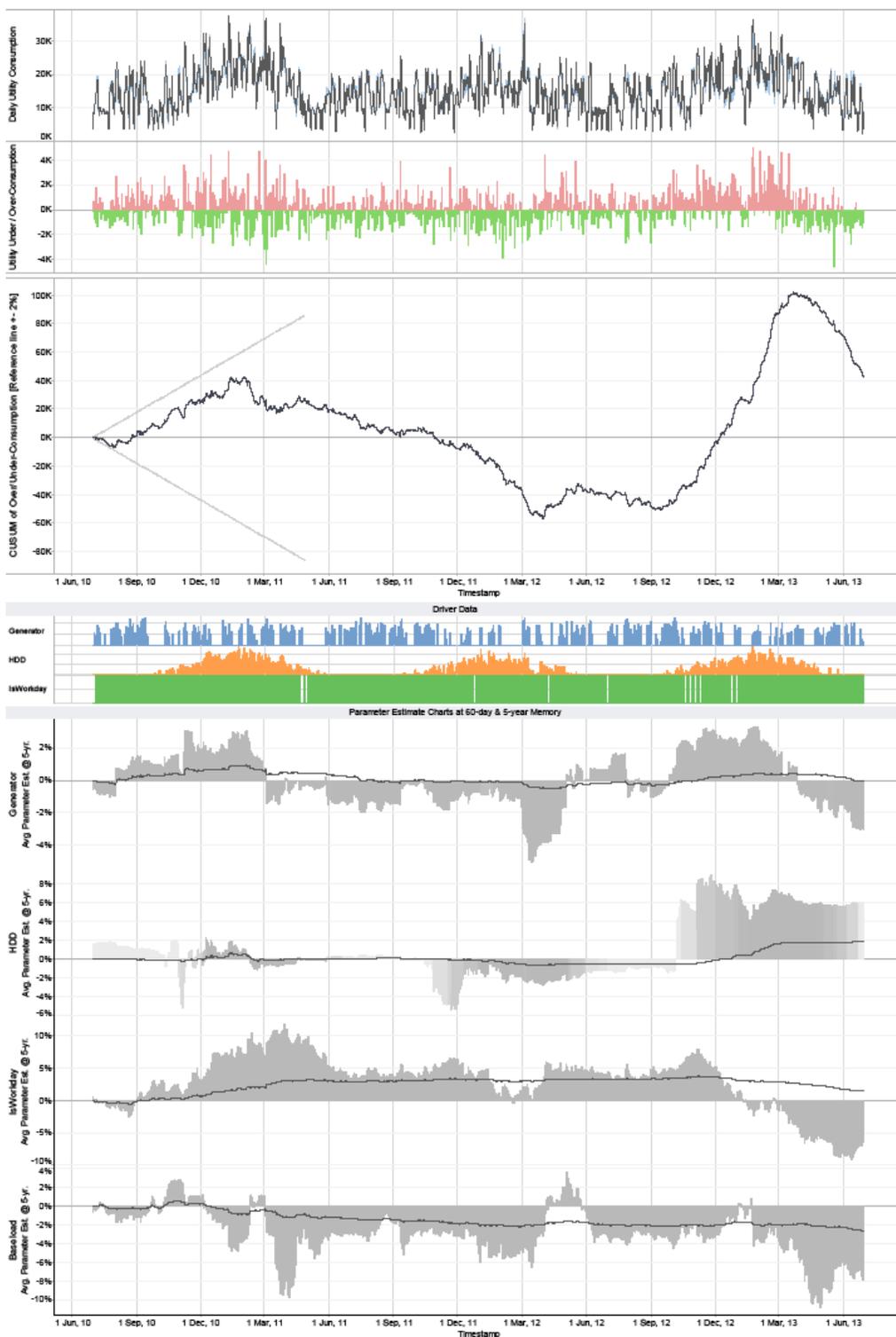
Scenario # 7 of 10



Scenario # 8 of 10



Scenario # 9 of 10



Scenario # 10 of 10

