CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

| | |
|---|---|
| Title | Development and evaluation of tools and methodologies for estimating behaviour and predicting training outcome of working dogs |
| Authors | Marcato, Marinara |
| Publication date | 2023 |
| Original Citation | Marcato, M. 2023. Development and evaluation of tools and methodologies for estimating behaviour and predicting training outcome of working dogs. PhD Thesis, University College Cork. |
| Type of publication | Doctoral thesis |
| Rights | © 2023, Marinara Marcato. - https://creativecommons.org/licenses/by-nc-nd/4.0/ |
| Download date | 2026-09-15 00:54:55 |
| Item downloaded from | https://hdl.handle.net/10468/15906 |

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na hÉireann, Corcaigh

**National University of Ireland, Cork**

# PhD Engineering Science

## Development and evaluation of tools and methodologies for estimating behaviour and predicting training outcome of working dogs

Thesis presented by

Marinara Marcato

0000-0002-7938-5227

for the degree of

**Doctor of Philosophy**

**University College Cork**

TYNDALL NATIONAL INSTITUTE

Head of School: Dr Fatima Gunning

Supervisors: Dr Paul Galvin
Dr Brendan O'Flynn
Dr Conor O'Mahony
Dr Salvatore Tedesco

2023

# Contents

List of Figures . . . . . v
List of Tables . . . . . x
Declaration of Responsibility . . . . . xiv
Acknowledgements . . . . . xvi
Abstract . . . . . xvii

**1 Introduction** **1**
1.1 Research Project . . . . . 1
1.2 Motivation . . . . . 1
1.3 Problem Statement . . . . . 2
1.4 Objectives . . . . . 3
1.5 Research Goals . . . . . 3
1.6 Novelty and Contribution . . . . . 4
1.7 Thesis Summary & Research Questions . . . . . 5
1.8 Research Contributions . . . . . 7

**2 Literature Review** **9**
2.1 Introduction . . . . . 9
2.2 Objectives . . . . . 10
2.3 Methodology . . . . . 11
2.4 Selection and Training Considerations . . . . . 12
2.4.1 Breed Considerations . . . . . 12
2.4.2 Genetic Considerations . . . . . 14
2.4.3 Training Considerations . . . . . 17
2.4.4 Discussion on Selection and Training Considerations 18
2.5 Behavioural Assessment Methods . . . . . 19
2.5.1 Test Batteries . . . . . 20
2.5.1.1 Puppy Testing . . . . . 21
2.5.1.2 Juvenile Testing . . . . . 29
2.5.1.3 Adult Testing . . . . . 30
2.5.2 Observational Tests . . . . . 33
2.5.3 Ratings . . . . . 34
2.5.3.1 Canine Personality Assessment . . . . . 35
2.5.3.2 Suitability for Assistance Work . . . . . 39
2.5.4 Mixed Assessment Methods . . . . . 41
2.5.5 Canine Activity Monitors . . . . . 42
2.5.6 Discussion on Behavioural Assessments . . . . . 50
2.6 Physiological Assessment Methods . . . . . 53
2.6.1 Systems . . . . . 54
2.6.2 Parameters . . . . . 55
2.6.2.1 Cardiac . . . . . 56
2.6.2.2 Respiratory . . . . . 58
2.6.3 Devices . . . . . 59
2.6.3.1 Polar H10 . . . . . 63

2.6.3.2 Bioharness . . . 64
2.6.4 Discussion on Physiological Assessments . . . 65
2.7 Discussion . . . 66

**3 Ratings** **68**
3.1 Introduction . . . 68
3.2 Methodology . . . 70
3.2.1 Ethical Approval . . . 70
3.2.2 Subjects . . . 70
3.2.3 Puppy Raiser Questionnaire - C-BARQ . . . 70
3.2.4 Dog Trainer Questionnaire - MCPQ-R . . . 71
3.2.5 Data Analysis . . . 72
3.3 Results . . . 77
3.3.1 Puppy Raiser Questionnaire - C-BARQ . . . 77
3.3.1.1 Feature Selection . . . 78
3.3.1.2 Model . . . 81
3.3.1.3 Performance . . . 82
3.3.1.4 Application . . . 83
3.3.1.5 Feature Importance . . . 84
3.3.2 Dog Trainer Questionnaire - MCPQ-R . . . 85
3.3.2.1 Feature Selection . . . 86
3.3.2.2 Model . . . 87
3.3.2.3 Performance . . . 88
3.3.2.4 Application . . . 88
3.3.2.5 Feature Importance . . . 90
3.4 Discussion . . . 91
3.4.1 Puppy Raiser Questionnaire - C-BARQ . . . 93
3.4.2 Dog Trainer Questionnaire - MCPQ-R . . . 94
3.4.3 Comparison between C-BARQ and MCPQ-R Models . . . 95
3.4.4 Comparison with Previous Research . . . 97
3.4.4.1 C-BARQ and Training outcome . . . 97
3.4.4.2 Prediction Performance . . . 101

**4 Assistance Dog Test Battery (ADBT)** **103**
4.1 Introduction . . . 103
4.2 Methodology . . . 105
4.2.1 Ethical Approval . . . 105
4.2.2 Subjects . . . 105
4.2.3 Test Battery . . . 105
4.2.4 Test Arena . . . 114
4.2.5 Data Analysis . . . 115
4.2.5.1 Validity . . . 116
4.2.5.2 Predictability . . . 116
4.3 Results . . . 119
4.3.1 Data Collection 1 . . . 119
4.3.1.1 Validity . . . 120
4.3.1.2 Predictability . . . 120

4.3.2 Data Collection 2 . . . 127
4.3.2.1 Validity . . . 128
4.3.2.2 Predictability . . . 129
4.4 Discussion . . . 137
4.4.1 Data Collection 1 . . . 138
4.4.2 Data Collection 2 . . . 138
4.4.3 Comparison between DC1 and DC2 Models . . . 139
4.4.4 Comparison with Previous Research . . . 143
4.4.4.1 Test Batteries and Training Outcome . . . 143
4.4.4.2 Prediction Performance . . . 146

**5 Posture Estimation** **150**
5.1 Introduction . . . 150
5.2 Methodology . . . 151
5.2.1 Subjects . . . 151
5.2.2 Devices and Data . . . 151
5.2.3 Data Collection Protocol . . . 152
5.2.4 Data Annotation . . . 152
5.2.5 Feature Extraction . . . 153
5.2.6 Classification Models . . . 154
5.2.6.1 Classifier 1 . . . 156
5.2.6.2 Classifier 2 . . . 156
5.2.6.3 Classifier 3 . . . 157
5.3 Results . . . 158
5.3.1 Classification Models . . . 158
5.3.1.1 Classifier 1 . . . 158
5.3.1.2 Classifier 2 . . . 158
5.3.1.3 Classifier 3 . . . 159
5.3.2 Feature Importance . . . 159
5.3.3 Intra-Breed Evaluation . . . 160
5.3.4 Inter-Breed Evaluation . . . 162
5.3.5 Best Classifier . . . 164
5.4 Discussion . . . 166
5.4.1 Comparison with Previous Research . . . 169
5.4.2 Limitations and Opportunities . . . 172

**6 Conclusion** **173**
6.1 Comparison between the Ratings and Test Battery . . . 174
6.2 Concluding Summaries . . . 175
6.3 Recommendations and Future Work . . . 179

**Bibliography** **185**

**Acronyms** **208**

**Appendix** **209**

**A Physiological Assessment - Hormonal** **210**

**B Behavioural Assessment - Ratings** **213**
B.1 Canine Behavioral Assessment and Research Questionnaire (C-BARQ) . . . 213
B.1.1 Descriptive Statistics . . . 218
B.1.2 Univariate Analysis . . . 226
B.1.3 Visualisation . . . 233
B.2 Monash Canine Personality Questionnaire - Revised (MCPQ-R) . . . 237
B.2.1 Descriptive Statistics . . . 237
B.2.2 Univariate Analysis . . . 238
B.2.3 Visualisation . . . 239

**C Behavioural Assessment - Test Battery** **244**
C.1 Ethogram . . . 244
C.2 Composite scores . . . 250
C.3 Data Collection 1 (DC1) . . . 253
C.3.1 Descriptive Statistics . . . 253
C.3.2 Univariate Analysis . . . 259
C.3.3 Visualisation . . . 266
C.4 Data Collection 2 (DC2) . . . 269
C.4.1 Descriptive Statistics . . . 269
C.4.2 Univariate Analysis . . . 275
C.4.3 Visualisation . . . 282

# List of Figures

2.1 Polar H10 heart rate sensor. . . . . 63
2.2 Bioharness, Zephyr . . . . . 64

3.1 Confusion matrix for the predicted and true outcome. . . . 74
3.2 Duration of training in weeks for dogs that were withdrawn from training. . . . . 78
3.3 Correlation matrix of variables considered for the C-BARQ reduced feature set. . . . . 81
3.4 Explained variance per principal component (Scree Plot). . 82
3.5 ROC and PR curves for the best logistic regression model using C-BARQ data selected by RFE. . . . . 83
3.6 Probability of withdrawal predicted by the best logistic regression model using C-BARQ data per true outcome (fail, success). . . . . 84
3.7 Probability of withdrawal predicted by the best logistic regression model using C-BARQ data per true status (withdrawn, service dog or guide dog). . . . . 84
3.8 PC01 coefficient loading from original items and factors. . . 85
3.9 Duration of training in weeks for dogs that were withdrawn from training. . . . . 86
3.10 Correlation Matrix for variables considered for the MCPQ-R reduced feature set. . . . . 87
3.11 Explained variance per principal component (Scree Plot). . 88
3.12 ROC and PR curves for the best logistic regression model using MCPQ-R data selected by RFE. . . . . 89
3.13 Probability of withdrawal predicted by the best logistic regression model using MCPQ-R data per true outcome (fail, success). . . . . 89
3.14 Probability of withdrawal predicted by the best logistic regression model using MCPQ-R data per true status (withdrawn, service dog or guide dog). . . . . 90
3.15 PC1 coefficient loading from original items. . . . . 91

4.1 Schematic representation of the test arena including key materials used and information about the test battery. . . . 114
4.2 Materials used in the test battery. . . . . 115
4.3 Duration of training in weeks for dogs that were withdrawn from training. . . . . 119
4.4 Correlation matrix for the numerical variables in Step 3 of feature selection. . . . . 122
4.5 ROC and PR curves for the optimal logistic regression model using ADTB data from DC1 selected by RFE. . . . . 126

4.6 Probability of failure predicted by the best logistic regression model using DC1 data per true working outcome (fail, success). . . . 127
4.7 Probability of failure predicted by the best logistic regression model using DC1 data per true status at the end of training (W = withdrawn, AD = assistance dog, and GD = guide dog). . . . 127
4.8 Duration of training in weeks for dogs that were withdrawn from training. . . . 128
4.9 Correlation matrix for the numerical variables in Step 3 of feature selection. . . . 131
4.10 ROC and PR curves for the optimal logistic regression model using ADTB data from DC2 selected by RFE. . . . 135
4.11 Probability of failure predicted by the best logistic regression model using DC2 data per true working outcome (fail, success). . . . 136
4.12 Probability of failure predicted by the best logistic regression model using DC2 data per true status at the end of training (W = withdrawn, AD = assistance dog, and GD = guide dog). . . . 136

5.1 IMUs sensor placement on the dogs' neck, back, and chest. . 152
5.2 Rolling window used for selecting data for feature extraction. Hyper-parameters shown include re-sampling transition time (t_time) between postures, window size (w_size) used to calculate the statistical measures and window offset (w_offset) to create separate observations. . . . 154
5.3 Direct estimator used in Classifier 1. . . . 156
5.4 Type, Static, and Dynamic estimators used in Classifier 2. . 157
5.5 Anomaly and Normal estimators used in Classifier 3. . . . 158
5.6 Confusion matrix for Classifier 1 on the test set. . . . 162
5.7 Confusion matrix for Classifier 2 on the test set. . . . 162
5.8 Confusion matrix for Classifier 3 on the test set. . . . 163
5.9 Confusion matrix for Classifier 1 on GR set. . . . 164
5.10 Confusion matrix for Classifier 2 on GR set. . . . 164
5.11 Confusion matrix for Classifier 3 on GR set. . . . 165
5.12 Confusion matrix for Classifier 3 on test and GR sets. . . . 166

B.1 Bar chart and box plot of the distribution of the C-BARQ item "I_01 When off the leash returns immediately when called" in relation to the working outcome. . . . 233
B.2 Bar chart and box plot of the distribution of the C-BARQ item "I_27 Toward cats squirrels or other small animals entering your yard" in relation to the working outcome. . . . 233
B.3 Bar chart and box plot of the distribution of the C-BARQ item "I_75 Chases or would chase birds given the opportunity" in relation to the working outcome. . . . 234

B.4 Bar chart and box plot of the distribution of the C-BARQ item "I_82 Begs persistently for food when people are eating" in relation to the working outcome. . . . . 234
B.5 Bar chart and box plot of the distribution of the C-BARQ item "I_85 Pulls excessively hard when on the leash" in relation to the working outcome. . . . . 234
B.6 Bar chart and box plot of the distribution of the C-BARQ item "I_90 Hyperactive restless has trouble settling down" in relation to the working outcome. . . . . 235
B.7 Bar chart and box plot of the distribution of the C-BARQ item "I_99 Licks people or objects excessively. . . . . 235
B.8 Bar chart and box plot of the distribution of the C-BARQ factor "F_02 Stranger directed aggression" in relation to the working outcome. . . . . 235
B.9 Bar chart and box plot of the distribution of the C-BARQ factor "F_06 Nonsocial fear" in relation to the working outcome. 236
B.10 Bar chart and box plot of the distribution of the C-BARQ factor "F_08 Dog directed fear." . . . . 236
B.11 Bar chart and box plot of the distribution of the C-BARQ factor "F_10 Separation related behaviour." . . . . 236
B.12 Bar chart and box plot of the distribution of the C-BARQ factor "F_11 Excitability." . . . . 237
B.13 Bar chart and box plot of the distribution of the MCPQ-R item "Motivation Persevering" in relation to training outcomes. . . . . 240
B.14 Bar chart and box plot of the distribution of the MCPQ-R item "Training Attentive" in relation to training outcomes. . 240
B.15 Bar chart and box plot of the distribution of the MCPQ-R item "Training Obedient" in relation to training outcomes. . 240
B.16 Bar chart and box plot of the distribution of the MCPQ-R item "Training Reliable" in relation to training outcomes. . 241
B.17 Bar chart and box plot of the distribution of the MCPQ-R item "Training Trainable" in relation to training outcomes. 241
B.18 Bar chart and box plot of the distribution of the MCPQ-R item "Amicability Easy-going" in relation to training outcomes. . . . . 241
B.19 Bar chart and box plot of the distribution of the MCPQ-R item "Amicability Friendly" in relation to training outcomes. 242
B.20 Bar chart and box plot of the distribution of the MCPQ-R item "Amicability Relaxed" in relation to training outcomes. 242
B.21 Bar chart and box plot of the distribution of the MCPQ-R item "Amicability Sociable" in relation to training outcomes. 242
B.22 Bar chart and box plot of the distribution of the MCPQ-R item "Neuroticism Nervous" in relation to training outcomes. 243

B.23 Bar chart and box plot of the distribution of the MCPQ-R item “Neuroticism Submissive” in relation to training outcomes. . . . 243
B.24 Bar chart and box plot of the distribution of the MCPQ-R item “Neuroticism Timid” in relation to training outcomes. 243

C.1 Bar chart and box plot of the distribution of the ADTB item “Tea Towel First Response Indifferent” at DC1 in relation to training outcomes. . . . 266
C.2 Bar chart and box plot of the distribution of the ADTB item “Distractions First Response - Human” at DC1 in relation to training outcomes. . . . 267
C.3 Bar chart and box plot of the distribution of the ADTB item “Crate Behaviours - Actively Seeking Attention” at DC1 in relation to training outcomes. . . . 267
C.4 Bar chart and box plot of the distribution of the ADTB score “S Distractions Pull_mean” at DC1 in relation to training outcomes. . . . 267
C.5 Bar chart and box plot of the distribution of the ADTB score “S Sensitivity” at DC1 in relation to training outcomes. . . . 268
C.6 Bar chart and box plot of the distribution of the ADTB score “S Crate Stimulus_mean” at DC1 in relation to training outcomes. . . . 268
C.7 Bar chart and box plot of the distribution of the ADTB score “S Petting Stimulus During_mean” at DC1 in relation to training outcomes. . . . 268
C.8 Bar chart and box plot of the distribution of the ADTB item “Tea Towel Second Response Indifferent” at DC2 in relation to training outcomes. . . . 282
C.9 Bar chart and box plot of the distribution of the ADTB item “Distractions First Response Car” at DC2 in relation to training outcomes. . . . 283
C.10 Bar chart and box plot of the distribution of the ADTB score “S Walking Pull_mean” at DC2 in relation to training outcomes. . . . 283
C.11 Bar chart and box plot of the distribution of the ADTB score “S Sensitivity” at DC2 in relation to training outcomes. . . . 283
C.12 Bar chart and box plot of the distribution of the ADTB score “S Kong Response_prod” at DC2 in relation to training outcomes. . . . 284
C.13 Bar chart and box plot of the distribution of the ADTB score “S Crate Handler_mean” at DC2 in relation to training outcomes. . . . 284
C.14 Bar chart and box plot of the distribution of the ADTB score “S Isolation Stimulus” at DC2 in relation to training outcomes. 284

C.15 Bar chart and box plot of the distribution of the ADTB score "S Isolation Handler_mean" at DC2 in relation to training outcomes. . . . . 285
C.16 Bar chart and box plot of the distribution of the ADTB score "S Familiarisation Handler_mean" at DC2 in relation to training outcomes. . . . . 285

# List of Tables

2.1 Test Batteries used for assessing apprentice assistance dogs. 23
2.2 Rating systems used to assess canine personality or suitability for assistance work. . . . . 36
2.3 Canine posture recognition algorithms using IMUs reported in literature. . . . . 44
2.4 Multi-parameter physiological devices and cardiac monitors physically composed of a sensing module attached to a chest strap. . . . . 61

3.1 C-BARQ rating sections, items and number of items per section. . . . . 71
3.2 MCPQ-R rating dimensions, items, and the number of items per section. . . . . 72
3.3 MCPQ-R rating items and items. . . . . 72
3.4 Variables with more than 5% missing values in the C-BARQ dataset. . . . . 79
3.5 Variables considered for the reduced feature set and the number of missing values. . . . . 80
3.6 Best logistic regression model using selected by RFE. . . . . 82
3.7 Best logistic regression model using selected by RFE. . . . . 88
3.8 C-BARQ factors significantly related to training outcome in assistance dogs. . . . . 99

4.1 Detailed description of the protocol for data collection using the ADTB. . . . . 106
4.2 Variables that satisfied the three-step feature selection strategy in DC1. . . . . 121
4.3 Optimal multivariate logistic regression model selected by RFE. . . . . 122
4.4 Univariate logistic regression model results of the variables in the optimal multivariate logistic regression model. . . . . 123
4.5 Composite scores, original ethogram items, and their respective marking for the variables chosen for the optimal model. . . . . 124
4.6 Importance of features in the optimal model. . . . . 128
4.7 Variables that satisfied the three-step feature selection strategy in DC2. . . . . 130
4.8 Optimal multivariate logistic regression model using selected by RFE. . . . . 131
4.9 Univariate logistic regression model results of the variables in the optimal multivariate logistic regression model. . . . . 132
4.10 Composite scores, original ethogram items, and their respective marking for the variables chosen for the optimal model. . . . . 133

4.11 Importance of features in the optimal model. . . . . 137
4.12 Ethogram items and composite scores are separated per subtest and by Data Collection 1 (DC1) and Data Collection 2 (DC2). Variables with a level of significance from univariate logistic regression below the threshold for retention ($p < 0.15$) are included. The variables highlighted in bold were selected for the optimal multivariate models. . . . . 140
4.13 Performance metrics from the models created with data from Data Collection 1 (DC1) and Data Collection 2 (DC2). . 142
4.14 Comparison between the performance results achieved by the two classifiers in this study and previously reported research. . . . . 147

5.1 Categories of behaviour analysed. . . . . 153
5.2 Type, posture, and number of observations in the IMU Posture dataset. . . . . 153
5.3 Number of observations after feature extraction per dataset per posture. . . . . 154
5.4 Grid search hyper-parameter set for the classifiers. . . . . 155
5.5 Classifier 2 model's hyper-parameter set selected by grid search and f1-weighted performance on 10-fold validation sets (mean $\pm$ standard deviation) and test set. . . . . 159
5.6 Classifier 3 model's hyper-parameter set selected by grid search and f1-weighted performance on 10-fold validation sets (mean $\pm$ standard deviation) and test set. . . . . 159
5.7 Feature scores calculated by f-classification in Select K Best by IMU Position and Sensor. . . . . 160
5.8 Feature importance calculated by Random Forest classifier considering the 80 features previously selected by Select K Best. . . . . 160
5.9 Feature importance calculated by Random Forest classifier considering the 80 features selected by Select K Best by domain. . . . . 161
5.10 Classification metrics per posture achieved using the best models selected by grid search in Classifier 1 on the test set. 161
5.11 Classification metrics per posture achieved using the best models selected by grid search in Classifier 2 on the test set. 162
5.12 Classification metrics per posture achieved using the best models selected by grid search in Classifier 3 on the test set. 163
5.13 Classification metrics per posture achieved using the best models selected by grid search in Classifier 1 on GR set. . . . 164
5.14 Classification metrics per posture achieved using the best models selected by grid search in Classifier 2 on GR set. . . . 165
5.15 Classification metrics per posture achieved using the best models selected by grid search in Classifier 3 on GR set. . . . 165

5.16 Classification metrics per posture achieved using the best models selected by grid search in Classifier 3 on test and GR sets combined. . . . . 166
5.17 Comparison between Classifier 3 performance and previous studies reporting inter-subject classification performance metrics per posture. The best geometric means between sensitivity and specificity (g-mean) are in bold. . . . 170

6.1 Performance metrics from the models created with data from the ratings C-BARQ and MCPQ-R and ADTB at DC1 and DC2. . . . . 174

B.1 Detailed description of the factors and the original items used to calculate them. . . . . 213
B.2 C-BARQ rating sections and items. . . . . 214
B.3 C-BARQ rating descriptive statistics including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing values for all items and factors. . . . . 219
B.4 Univariate logistic regression model estimates, p-value, Odds Ratio (OR), and Confidence Interval (CI). . . . . 226
B.5 MCPQ-R rating sub-scales and items. . . . . 237
B.6 MCPQ-R rating descriptive statistics including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing values for all items and factors. . . . . 238
B.7 Univariate logistic regression model estimates, p-value, Odds Ratio (OR), and Confidence Interval (CI). . . . . 239

C.1 Detailed description of the ethogram for each of the subtests in the novel Assistance Dog Test Battery (ADTB), including the behavioural items analysed and the marking system. 244
C.2 Composite scores originated from Assistance Dog Test Battery (ADTB) ethogram items presented in Table C.1. For numerical ethogram items, composite scores were calculated by taking the mean and product between them. For categorical ethogram items, composite scores were derived by assigning numbers to categories in a meaningful way. . . . 250
C.3 Descriptive statistics for DC2 numerical variables of the ADTB ethogram including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing values for all items and scores. . . . . 254
C.4 Descriptive statistics of DC1 for categorical variables of the ADTB ethogram including the number of No, Yes and missing data. . . . . 258
C.5 ADTB ethogram descriptive statistics for the variable Kong Interaction Lateralisation, including the number of dogs in the sample per category. . . . . 259

C.6 Descriptive statistics of the composite scores used in the optimal multivariate logistic regression model and the items from which they derive. . . . . 259
C.7 ADTB univariate logistic regression model estimates, p-value, Odds Ratio (OR), and Confidence Interval (CI). . . . . 261
C.8 Univariate logistic regression model using the demographic variable Breed as a predictor of the probability of withdrawal, including estimates, p-value, Odds Ratio (OR) and Confidence Interval (CI). . . . . 266
C.9 Univariate logistic regression model using the variable Kong Interaction Lateralisation as a predictor of the probability of withdrawal. Model parameters reported include estimates, p-value, Odds Ratio (OR) and Confidence Interval (CI). . . . . 266
C.10 Descriptive statistics for DC2 numerical variables of the ADTB ethogram including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing values for all items and scores. . . . . 270
C.11 Descriptive statistics of DC2 for categorical variables of the ADTB ethogram including the number of No, Yes and missing data. . . . . 274
C.12 ADTB ethogram descriptive statistics for the variable Kong Interaction Lateralisation, including the number of dogs in the sample per category. . . . . 275
C.13 Descriptive statistics of the features used in the optimal multivariate logistic regression model and the items from which they derive. . . . . 275
C.14 ADTB univariate logistic regression model estimates, p-value, Odds Ratio (OR), and Confidence Interval (CI). . . . . 277
C.15 Univariate logistic regression model using the demographic variable Breed as a predictor of the probability of withdrawal, including estimates, p-value, Odds Ratio (OR) and Confidence Interval (CI). . . . . 282
C.16 Univariate logistic regression model using the variable Kong Interaction Lateralisation as a predictor of the probability of withdrawal. Model parameters reported include estimates, p-value, Odds Ratio (OR) and Confidence Interval (CI). . . . . 282

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism.

*Marinara Marcato*

# Dedication

First and foremost, I would like to thank my parents Gilberto and Maria for their unconditional love and support. For they have taught me the most important lessons and skills. The more I experience in life the more deeply I understand the fundamental role they have in my life. I am forever grateful for your life. I love you to the infinity.

To my brothers, Felipe and Lucas, thank you for all the lessons you taught me and all the joy that you bring into my life. To my nephews and niece, Heitor, Henrique, and Heloisa; you gave the words happiness and love the purest and most profound meaning. My sincere gratitude to everyone in my family for all that you represent to me.

To my boyfriend, Fionn, for your continued support, believing in me and cheering me through the ups and downs during all these years. To my friends, you know who you are, for making my days more fun, bringing balance to my life, and putting my challenges in perspective. In particular, to Caroline and Larissa, for sharing this journey and home with me in Cork.

To my teachers and professors, for not only teaching me but also inspiring me to continue learning and showing me opportunities for future endeavours. I am grateful to all those who contributed to my education from primary, through secondary to graduate level.

I certainly would not have gotten here without your support; therefore, this achievement is also yours.

# Acknowledgements

I would like to take this opportunity to thank my supervisors, Paul, Brendan, and Connor, for believing in my potential and supporting me along this journey. Salvatore offered me excellent support during the last couple of years. Your advice and guidance were fundamental to the successful completion of this degree.

To everyone in Tyndall for providing an ideal ecosystem for me to pursue my PhD. I appreciate having had the opportunity to connect with people from all over the world and learn with and from them in such an inclusive environment. Your friendship has made my journey very enjoyable.

I would like to thank Maria Vittoria Lorenzi, Hazel Craven, Conghal Hewson, and Jennifer Kenny, for their contribution to the data collection and the opportunity to co-supervise their work.

I am grateful to the Irish Guide Dogs for the Blind and Swansea University for their partnership. In particular, to all staff members and volunteers who contributed to this research project during the design and data collection phases.

Lastly, this research was made possible through the financial support of the following intuitions: The Ireland-Wales INTERREG Programme under the CALIN project funded by the European Regional Development Fund (ERDF) (grant number 80885), from Science Foundation Ireland (SFI) (grant number 12/RC/2289-P2), and SFI and Dept of Agriculture, Food and Marine (grant Number 16/RC/3835).

# Abstract

**Background:** The average training success rate in different dog industries is as low as 50% and the cost of training a guide dog is as high as €53,00 in Ireland. The key to reducing costs is in the assessment of trainee dogs for identifying likely to fail at an early stage.
**Objectives:** This thesis aims to improve behavioural assessment methods by including machine learning methods to (1) predict future outcomes in trainee assistance dogs based on ratings and test batteries, and (2) estimate canine posture based on a recognition system specifically designed for working dogs.
**Methods:** (1) Two standardised ratings were used, in particular, the Canine Behavioral Assessment and Research Questionnaire (C-BARQ) was completed by puppy raisers and the Monash Canine Personality Questionnaire - Revised (MCPQ-R) was answered by dog trainers. Rating data were independently analysed to investigate their relationship with training outcomes. The novel Assistance Dog Test Battery (ADTB) was designed to assess the suitability of trainee assistance dogs for assistance work during training. The test was conducted at 3 weeks - Data Collection 1 (DC1) - and 10 weeks - Data Collection 2 (DC2) - after the start of formal training to investigate the optimal timing to predict working outcomes. (2) Three Inertial Measurement Units (IMUs) were placed on the dogs in different positions (neck, back and chest) and five postures (walking, standing, sitting, lying down and body shake) were annotated. Advanced machine learning techniques were applied for the first time in this field to improve state-of-the-art posture prediction performance.
**Results:** (1) The machine learning models achieved an area under the ROC of 0.84 and 0.85 when using the ratings C-BARQ and MCPQ-R to predict training outcome; and 0.74 and 0.84 when using the DC1 and DC2 of the ADTB to predict working outcomes, respectively. (2) The optimal canine posture classifier achieved an f1-weighted of 0.90.
**Conclusions:** (1) These novel machine learning models provided the most effective early prediction of suitability for assistance work. The MCPQ-R and ADTB were demonstrated for the first time to be a reliable canine behavioural assessment method for estimating future outcomes in trainee dogs. (2) Comparison with previous work reveals a superior performance of the new canine posture estimation system for working

dogs. Enhanced feature engineering can improve performance of machine learning models for predictive methods associated with working dog training. This research contributes to the advancement of the state-of-the-art by supplying datasets containing behavioural, inertial and physiological data following FAIR data guidelines.

# Chapter 1

# Introduction

## 1.1 Research Project

This PhD formed part of the Guide Dog 4.0 Project, which is a collaborative project between Tyndall National Institute (TNI), Swansea University (SU), and Irish Guide Dogs for the Blind (IGDB). These institutions came together under the Celtic Advanced Life Science Innovation Network (CALIN), as part of the Ireland-Wales INTERREG Programme, with a common objective of improving guide and assistance dog training programmes.

The collaboration leverages the expertise available in each institution in the areas of wearable technology and data science at TNI, physiological biomarker profiling at SU, and dog training and behaviour at IGDB. This PhD research project aims to further advance state-of-the-art knowledge and practices related to assessing the suitability of apprentice dogs for the assistance training programme.

## 1.2 Motivation

Humans rely on working dogs in diverse domains, for instance, guide dogs aid visually impaired people, search and rescue dogs can locate people, detection dogs can recognise substances and medical dogs can identify certain blood sugar levels and cancer. Working dogs have been trained in such roles for a variety of economic reasons or their natural

aptitude. In some cases, it is just less costly or complicated for a dog to do the job, while in others machines or humans are not as capable.

Valuable resources are dedicated to supplying these highly trained canines. For instance, prospective assistance dogs at the Irish Guide Dogs for the Blind (IGDB) start training at the age of 8 weeks and join the other dogs at their Training Centre at 14 months old, qualifying at 20-24 months old (IGDB 2018). During this period, the dog's temperament and ability to perform the required tasks are assessed while positive behaviour and guiding skills are reinforced (Ziv 2017).

However, some dogs turn out to be unsuitable for guide dog training as described in the following paragraphs. Previous research discloses an average failure rate as high as 50% in different dog industries (Sinn et al. 2010; Wilsson et al. 1997b) (Asher et al. 2013; L. M. Tomkins et al. 2011). As a result, guide dogs can cost €53,000 to produce in Ireland (IGDB 2021) and the annual cost associated with ownership averages almost €5,000.

Therefore, the early identification of suitable dogs brings several benefits. Firstly, the cost of production can be reduced by early re-homing of unsuitable dogs which also allows the saved resources to be used on dogs with more significant potential. Secondly, the levels of welfare can be improved by early withdrawing unsuitable dogs from training, as they could perceive training as a challenging experience due to their innate unsuitability (L. M. Tomkins et al. 2011).

## 1.3 Problem Statement

This section briefly introduces the problem statements associated with this body of work for the purpose of providing context to the next sections. These problems will be discussed in more detail in Chapter 2. Two main practical problems related to the assessment of assistance dogs were identified:

1. The lack of a standardised method to accurately estimate the suitability of a prospective dog at an early stage of training hinders the ability to produce assistance dogs in the most timely and cost-effective manner;

**2.** The lack of an accurate canine activity recognition system for working dogs negatively impacts the ability to automatically detect and remotely monitor the activities of these highly trained canines.

## 1.4 Objectives

This PhD research thesis comprises two main objectives. The primary objective is to evaluate and compare the validity of traditional and novel behaviour assessment methods to predict training outcomes and working outcomes for trainee assistance dogs. The secondary objective is to improve the performance of canine posture recognition systems for working dogs by exploring the use of multiple sensors, data pre-processing techniques, and machine learning methods.

## 1.5 Research Goals

In order to achieve the two objectives outlined above, the following research goals were pursued:

**1.a.** Assess the validity of a standardised canine personality rating to predict training outcomes in prospective assistance dogs for the first time.

**1.b.** Compare the performance of a widely used behaviour rating for training outcome prediction and a personality rating never used for that purpose before.

**1.c.** Design and validate a novel test battery leveraging state-of-the-art research on working dog assessment with the goal of evaluating the working outcome.

**1.d.** Develop a new machine learning algorithm capable of predicting working outcomes in prospective assistance dogs using data from the novel test battery.

**1.e.** Investigate the best time during assistance dog training to perform the novel test battery for optimal working outcome predictions.

**1.f.** Compare the predictive power of the two standardised ratings and ethogram scoring of the novel test battery.

**2.a.** Evaluate the use of advanced time series feature extraction methods in a canine posture recognition system for the first time.

**2.b.** Investigate the predictive power of multiple inertial sensors considering their type and location to accurately predict key canine postures.

## 1.6 Novelty and Contribution

The novelty and contribution to the body of knowledge resulting from the present thesis are introduced below in order of appearance:

**Chapter 3** presents a novel application of MCPQ-R rating for training outcome prediction in prospective assistance dog. The model built using the MCPQ-R achieved performance results comparable to those of the widely used C-BARQ in the same sample of dogs. The main advantages of using MCPQ-R are due to the fact that it is considerably shorter than the C-BARQ (26 vs 100 items) and respondents are not required to be as familiar with the dog. Hence, the MCPQ-R was proven to be an appropriate rating to estimate the suitability of prospective assistance dogs.

**Chapter 4** describes the design of a novel test battery (ADTB) for assessing the suitability of trainee dogs for assistance work. Subsequently, it discusses the development of a machine learning algorithm capable of predicting working outcomes in trainee assistance dogs using the ethogram evaluations of the ADTB at three and ten weeks of training. The results achieved by the model built for predicting working outcomes were comparable to the widely used C-BARQ for predicting training outcomes and outperformed models built in previous studies. Hence, ADTB was proven to be an appropriate test battery to estimate the suitability of prospective assistance dogs.

This research originates the first dataset for the evaluation of prospective dogs' suitability for assistance work. It contains multi-parameter behavioural and physiological data of over 100 dogs. The following parameters were recorded and will be made publicly available: questionnaire data - ratings Canine Behavioral

Assessment and Research Questionnaire (C-BARQ) and Monash Canine Personality Questionnaire - Revised (MCPQ-R) -, behavioural data - ethogram markings from Assistance Dog Test Battery (ADTB); and inertial sensors Accelerometer and Gyroscope - and physiological data - cardiac (Interbeat Interval, Heart Rate Variability, Heart Rate), respiratory (Breathing Rate) and hormonal (cortisol).

**Chapter 5.** Development of the first canine posture recognition algorithm employing advanced feature extraction techniques for time series on inertial data. Sensor and feature type importance were analysed for the first time. The performance results revealed a superior performance compared to previous research by employing more IMUs and novel advanced machine learning techniques.

Moreover, the dog posture recognition dataset built specifically using the most common working dog breeds will be made publicly available. It contains inertial data (Accelerometer, Gyroscope and Magnetometer) collected from IMUs (back, neck, and chest) on 42 healthy trainee assistance dogs. The data were annotated considering the occurrence of five postures (lying down, sitting, standing, walking, and body shaking).

## 1.7 Thesis Summary & Research Questions

This thesis is divided into five chapters associated with the work carried out to achieve the research goals as detailed below. The research questions and summary of each chapter are presented below:

**Chapter 2 - Literature Review** introduces established practices in assistance dog training including selection, training, and behavioural assessment methods as well as experimental physiological assessment methods under investigation by the scientific community.

**Chapter 3 - Ratings** relates to the research goal 1.a. It uses two standardised questionnaires, in particular Canine Behavioral Assessment and Research Questionnaire (C-BARQ) (J. A. Serpell et al.

2001; Hsu et al. 2003; Duffy et al. 2012) and Monash Canine Personality Questionnaire - Revised (MCPQ-R) (Ley et al. 2009a; Ley et al. 2009a; Ley et al. 2009b). Puppy raisers answered the C-BARQ around the time the dogs started their formal training and dog trainers completed the MCPQ-R 10 weeks later. The resulting datasets are analysed to investigate the relationship between each variable and training outcome. Subsequently, the selected variables are fitted to a machine learning model for predicting the probability of being withdrawn from the training programme.

Research questions addressed:

Q.1 Is there a significant association between C-BARQ and MCPQ-R and training outcomes?

Q.2 Can statistical models estimate training outcomes based on C-BARQ and MCPQ-R ratings?

Q.3 Which is more predictive of training outcomes: the C-BARQ or the MCPQ-R?

**Chapter 4 - Assistance Dog Test Battery (ADTB)** relates to the research goals 1.b, 1.c, and 1.d. It presents the test battery protocol and ethogram rating system developed as the main novelty and component of this PhD project. It contains a detailed description of the methodology utilised including procedures, materials, devices and parameters recorded as part of Data Collection 1 (DC1) and Data Collection 2 (DC2) at 3 and 10 weeks respectively after the dogs started training. The resulting datasets are analysed to investigate the relationship between each variable and working outcomes. Subsequently, the selected variables are fitted to a machine learning model for predicting working outcomes.

Research questions addressed:

Q.4 Are the new ethogram items and composite scores developed for evaluating the novel ADTB valid for assessing working outcomes?

Q.5 Can machine learning models predict working outcomes using ethogram items and composite scores for evaluating the novel ADTB?

Q.6 When is the optimal time to perform the novel ADTB on trainee dogs to maximise working outcome prediction performance?

**Chapter 5 - Posture Estimation** relates to the research goals 2.a and 2.b. It presents the data collection methods and machine learning techniques employed to estimate canine posture based on IMU sensors. It aims at outperforming previous research on canine posture estimation systems by designing a system for working dogs using multiple inertial sensors and advanced machine learning techniques.

Research questions addressed:

Q.7 Which IMU sensor position yields the most predictive features for canine posture recognition?

Q.8 Which feature domain produces the most predictive features for canine posture recognition?

**Chapter 6 - Conclusion** relates to the research goal 1.e. It reviews the objectives, summarises the main findings and compares the results from Chapter 3 - Ratings and Chapter 4 - Assistance Dog Test Battery (ADTB). It presents recommendations relevant to practitioners and researchers, highlighting the potential for practical use and areas for future research.

Research questions addressed:

Q.9 Which of the behavioural assessment methods are more predictive of future outcome - the rating (C-BARQ or MCPQ-R datasets) or ADTB ethograms (DC1 or DC2 datasets)?

The **Appendix** contains a full description of the standardised ratings; namely, C-BARQ and MCPQ-R, used in Chapter 3, as well as relevant plots and data analysis results. Additionally, it contains a detailed presentation of the novel Assistance Dog Test Battery presented in Chapter 4 including the subtest protocol, ethogram scoring system, relevant plots, and data analysis results.

## 1.8 Research Contributions

Assistance dog selection and performance assessment methods using

behavioural and physiological tools and devices. Marcato M, Tedesco S, O'Mahony C, O'Flynn B, Galvin P (2022). Applied Animal Behaviour Science, Elsevier.

Comparison between the Canine Behaviour and Research Questionnaire and Monash Canine Personality Questionnaire - Revised to predict training outcome in apprentice assistance dogs. Applied Animal Behaviour Science Journal, Elsevier.

Prediction of working outcomes in trainee dogs using the novel Assistance Dog Test Battery (ADTB). Under review Applied Animal Behaviour Science Journal, Elsevier. Submitted in September 2023.

Machine learning based canine posture estimation using inertial data. Marcato M, Tedesco S, O'Mahony C, O'Flynn B, Galvin P (2023). PLOS ONE.

# Chapter 2

# Literature Review

## 2.1 Introduction

Humans rely on working dogs in a range of diverse domains. For instance, search and rescue dogs can locate missing people (K. E. Jones et al. 2004); police detection dogs can recognise certain substances such as explosives and narcotics (Jezierski et al. 2014); and assistance dogs provide support to individuals with physical, medical, or emotional disabilities (Davis et al. 2004). Working dogs have been deployed in these specific roles due to economic reasons or their natural aptitude for the task at hand. In some cases, it is less costly or complicated for a dog to do the job (Kokocińska-Kusiak et al. 2021), while in others machines or humans are simply not able to accomplish the task (Wiggett 2006).

The present literature review covers methods for selecting and assessing performance in prospective assistance dogs, with a particular focus on guide dogs. Three types of assistance dogs are defined by Assistance Dogs International (ADI 2022), namely guide dogs assist individuals with visual impairments, hearing dogs assist individuals with hearing impairments, and service dogs assist individuals with disabilities other than blindness or deafness. More specifically, as defined by Rodriguez et al. 2020, “service dogs are those who assist individuals with physical disabilities (e.g., performing mobility-related tasks such as pulling a wheelchair or retrieving dropped items), individuals with medical conditions (e.g., alerting or responding to medical crises such hypoglycemia or seizures (Wells et al. 2016)), and individuals with mental health disorders (e.g., psychiatric service dogs for post-traumatic stress disorder or autism spectrum disorder)”.

Valuable resources are dedicated to supplying these highly trained canines. Assistance dog programmes generally implement genetic selection and breeding, puppy rearing, early socialisation, recruitment and assessment, early and advanced training, and handler education

(Mia Cobb et al. 2015). In order to achieve the successful qualification, these dogs must respond appropriately to environmental stressors (large crowds, loud noises) and distractions (other animals and people, food on the ground), and exhibit impulse control, flexible and sustained attention, appropriate social behaviour, and independent problem-solving. Trainee dogs' temperament and behavioural suitability to perform the required tasks are continuously assessed while positive behaviour and desired skills are reinforced (Ziv 2017).

Data received from guide and assistance dog organisations have disclosed that programme failure rates of 50-70% are typical with behavioural reasons being the primary cause for 63-87% of those failures (Duffy et al. 2012). Extensive resources are invested in breeding, raising, and training these canines; as a result, the production of a guide dog can cost over €37,000 (Wirth et al. 2008). In Ireland, it can cost up to €53,000 to create a guide dog (IGDB 2021).

The accurate evaluation of proficiency in fundamental skills and prediction of successful qualification as early in the training programme as possible is crucial to increasing efficiency and reducing costs (Bray et al. 2019). More specifically, the three main benefits associated with the early identification of unsuitable dogs are as follows. Firstly, it reduces production costs as unsuitable dogs can be re-homed earlier. Secondly, it increases the withdrawn dogs' welfare as training may be perceived as a stressful experience by them (Arhant et al. 2020). Thirdly, it improves efficiency as trainers can focus on dogs with greater potential.

Therefore, a range of dog selection and assessment methods have been investigated, developed and employed by researchers and working dog organisations around the world to identify key behavioural traits that either lead to the success or failure of candidate dogs within the programme. Researchers have investigated the behavioural reasons for withdrawal from assistance dog training programmes and found factors related to (1) stress responses such as fear, aggression, nervousness, and anxiety (M. Goddard et al. 1983; Harvey et al. 2016b; Harvey et al. 2017; Duffy et al. 2012; Dollion et al. 2019; Bray et al. 2019); (2) distractibility or excitability - inability or unwillingness to concentrate, or slow to recover once distracted (M. E. Goddard et al. 1982; J. A. Serpell et al. 2001; Harvey et al. 2016b; Harvey et al. 2017; Arata et al. 2010); (3) body sensitivity (Harvey et al. 2017) - which indicates that they would not accept the harness.

## 2.2 Objectives

This thesis chapter discusses current practices and emerging research in selecting and assessing trainee dogs' suitability for assistance work. It

reviews relevant research in the following areas:

- Selection and training methods in the assistance dog industry: important considerations and common practices among organisations;
- Behaviour assessment methods: test batteries, ratings, observational tests, and activity monitors;
- Physiological assessment methods: physiological basis, relevant measures, Commercial Off-The-Shelf (COTS) devices;

Each of the three sections discusses the key findings, highlights patterns and trends, and presents recommendations for assistance dog organisations and further research needed to advance the state-of-the-art knowledge.

## 2.3 Methodology

Google Scholar and Science Direct were the search engines used to identify published papers. A combination of the following keywords was used where the brackets group similar keywords: (working dog, guide dog, assistance dog, service dog, dog), (behaviour test, temperament test, test battery, questionnaire, ratings, personality assessment), (performance assessment, behavioural assessment, training outcome), (canine activity monitors, canine activity trackers), genetics, (emotion, stress, emotional state, affective state), (physiological, biomarkers, heart rate, heart rate variability, respiratory rate, cortisol, sIgA, CgA). Some papers were also identified through the citations and references contained in other relevant articles. Google was the main search engine used to find canine activity monitors and physiological sensing devices. Some devices were found via papers where they were mentioned. The searches were performed between September/2020 and December/2021. The criteria used for selecting studies are discussed below for each of the sections.

Selection and Training Consideration (Section 2.4):

- Breed considerations (Section 2.4.1): It includes widely used breeds in assistance dog programmes. The most popular breeds were discussed in detail - namely, Golden and Labrador Retrievers, and German Shepherds - based on the current literature. The less popular breeds were mentioned; however, no publications on their capabilities as assistance dogs were found to allow a more extensive analysis.
- Genetic considerations (Section 2.4.2): Purpose-bred and non-purpose-bred assistance dogs were included in the analysis. The heritability of important behaviour traits and genetic disorders

(congenital heart disease, hip dysplasia) in assistance dogs were discussed based on the available literature.

- Training considerations (Section 2.4.3): This section includes a complete overview of the training programme. Published literature that discussed the training programme of assistance dogs was included in this section.

Behavioural Assessment Methods (Section 2.5):

- Test batteries (Section 2.5.1): Published in the literature in relation to assistance dogs. Details of sub-tests and variables scored are available, as well as tests suitable for different age groups (puppy, juvenile, adult).
- Observational tests (Section 2.5.2): Published in the literature and conducted in non-controlled environments that allowed for the surveillance of a dog's natural behaviour. Published in the literature in relation to assistance dogs.
- Ratings (Section 2.5.3): Published in the literature in relation to assessing the temperament of assistance dogs.
- Canine Activity Monitors (Section 2.5.5): Published in the literature relating to the development of a canine posture estimation system based on inertial sensors.

Physiological Assessment Methods (Section 2.6):

- Parameters (Section 2.6.2): Due to the limited number of studies involving physiological assessment of prospective assistance dogs with the goal of testing association with training outcome, the inclusion criteria were broadened. This section included papers relating to factors that were shown to affect the probability of completion of training such as emotional response, stress, trainability, etc, and their relationship with diverse physiological biomarkers in dogs or other animals.

## 2.4 Selection and Training Considerations

### 2.4.1 Breed Considerations

Throughout the world, there are approximately 400 recognised dog breeds, each with its own unique history, genetic profile, and reputation (van Rooy et al. 2019). Assistance dog facilities typically use the same limited number of breeds in their programs. These are Labrador Retrievers (LR), Golden Retrievers (GR), and a mix between the two breeds. Some also use German Shepherds (GS) (Ennik et al. 2006), Labradoodles, and Goldendoodles (Shouldice et al. 2019). Some training

facilities have also expanded their selection to include Royal Poodles, Bouvier des Flandres, Bernese Mountain Dog, Labrenese (LR x Bernese Mountain Dogs), and Saint-Pierre (Dollion et al. 2019). There are differences between breeds which can make some more suited to a role than others (Ennik et al. 2006). Geoffrey Caron-Lormier et al. 2016 conducted a study using guide dogs to investigate the patterns of undesirable behaviour in dogs. A dataset collected by Guide Dog UK staff over a 20-year period provided this study with information on 7770 dogs, where 6465 dogs reached retirement and 1305 dogs were withdrawn before retirement for behavioural reasons. Withdrawal decisions were made by the dog care and welfare manager.

In Geoffrey Caron-Lormier et al. 2016 it was found that Labrador Retrievers (N = 2877) are the most likely breed to reach retirement (i.e., the dog successfully completed their training programme, was partnered, and worked until retirement). They scored higher for nerve stability and were more cooperative when compared with German Shepherds. They were also the least fearful of the breeds assessed (Geoffrey Caron-Lormier et al. 2016). In Ennik et al. 2006 which assesses the suitability of four dog breeds for guide work, Labrador Retrievers (N = 690) scored higher in traits such as courage, hardness, and friendliness than other breeds (Ennik et al. 2006). van Rooy et al. 2019 focused on the behavioural genetics of dog breeds and found they were also ranked highly around the world as having low levels of aggression and scoring highest in trainability.

Golden Retrievers (N = 848) were also a chosen breed in Geoffrey Caron-Lormier et al. 2016 and were less likely to be withdrawn early due to reasons of environmental anxiety. They were also less likely to be withdrawn for social behaviour reasons (Geoffrey Caron-Lormier et al. 2016). van Rooy et al. 2019 found they have lower owner-directed aggression when compared to other breeds and exhibit low aggression levels towards unfamiliar humans and dogs. They also score high for trainability, but not as high as Labrador Retrievers. While working with Golden Retrievers (N = 407), Ennik et al. 2006 suggested that they may benefit from a prolonged training period. Golden Retrievers worked 36 days shorter than the average 3108 days (Geoffrey Caron-Lormier et al. 2016).

Ennik et al. 2006 found that Labrador Retrievers crossed with Golden Retrievers (N = 360) had the highest probability of becoming a guide dog, suggesting that the genetic mix between these two is favourable. The combination of both the gentleness of the Golden Retriever and the willingness of the Labrador Retriever allows this crossbreed to be a successful assistance dog (K. M. Evans et al. 2015).
Geoffrey Caron-Lormier et al. 2016 used a variety of Labrador Retrievers crossed with Golden Retrievers. They were grouped depending on generation, purebred cross and breeds of the sire and

dam. Golden Retriever x Labrador Retriever (N = 2167) scored best under the excitability group and were least likely to be withdrawn for that reason. However, second generation Labrador Retrievers crossed with a purebred Golden Retriever (N = 268) were less likely to reach retirement due to undesirable behaviour (not specified in the text) being displayed to the Guide Dog Owner. However, it has been suggested that older dogs often develop undesirable behaviours, such as increased distractibility, this, therefore, contributes to a breakdown in obedience.

German Shepherds have a strong instinct to control the movements of those around them. Their intelligence, trainability, and high activity levels make them good candidates for an assistance dog position (van Rooy et al. 2019). Ennik et al. 2006 found they (N = 576) often scored highest in traits such as sharpness and defence drive. In Geoffrey Caron-Lormier et al. 2016, German Shepherds (N = 286) were the most fearful of breeds assessed and were seven times more likely to be withdrawn for fear and aggression than the Labrador Retriever. They scored well in willingness and confidence, and are 70% less likely than the other breeds to be withdrawn because of these issues. They were less likely to reach retirement (3108 days) due to behavioural issues such as anxiety, fear, and aggression, and were withdrawn an average of 172 days early (Geoffrey Caron-Lormier et al. 2016). They are susceptible to several health conditions including musculoskeletal conditions, skin conditions, and nervous sensory conditions which shorten their working life (G. Caron-Lormier et al. 2016).

In recent years, Labradoodles, which is a cross between a Labrador Retriever and a Poodle, have become a popular breed of assistance dog. They are a purpose crossbreed originally intentionally created by Wally Conran to create a hypoallergenic assistance dog. Goldendoodles have also become a popular assistance dog breed. These are a Golden Retriever crossed with a Poodle (Shouldice et al. 2019). Current research indicates that Labradoodles and Goldendoodles are suitable for use in roles such as therapy dogs for people suffering from dementia (Zon 2017) and as diabetes detection dogs (Petry et al. 2015). However, there are no studies available which assess their capabilities as assistance dogs or their behaviour during training.

### 2.4.2 Genetic Considerations

Dogs used in assistance dog organisations are bred for the particular traits that are required in the execution of assistance work. These traits include intelligence, good health, and suitable temperament (Ennik et al. 2006). The success of any breeding programme involving behavioural traits in dogs depends greatly on the understanding of the genetics of such behavioural traits. The understanding of behavioural traits and their heritability could be further improved if data relating to

genes affecting canine behaviour was collected (Hradecká et al. 2015).

Breeding programs are advantageous as they offer the ability to choose dogs on the basis of heritable health and behavioural traits. To maintain diversity within these breeding programs, organisations often take dogs from other organisations (Ennik et al. 2006). Not only does this maintain diversity but it also gives the opportunity to genetically improve their existing stock (K. Evans et al. 2013).

Assistance dog breeds are susceptible to common genetic disorders. For example, congenital heart disease often affects Golden Retrievers and Labrador Retrievers from birth, and causes a variety of cardiac malformations of the heart and its vessels (Parenti et al. 2015). Studies have reported the successful implementation of genetic selection methods in breeding programmes in order to decrease the occurrence of inheritable genetic disorders over the years (Hedhammar 2020).

A popular example of a genetic disorder is hip dysplasia - a condition affecting the hip joint in which the joint develops incorrectly leading to inflammation and pain (Mackenzie et al. 1985). This condition affects all breeds used for assistance work; namely, Golden Retrievers, Labrador Retrievers, and German Shepherds. A study by Oberbauer et al. 2017 investigated the success of long-term genetic selection to prevent hip and elbow dysplasia in 60 dog breeds. The result showed that while effective breeding programmes can improve joint health, there is variability in heritability between breeds. Extensive research over decades has shown the value of calculating the genetic risk associated with each individual in the pedigree. The Swedish Kennel Club started implementing a metric called estimated breeding values (EBV) which is calculated considering pedigree information with hip status (Hedhammar 2020). They reported a reduction by one-third of hip dysplasia in their dogs through applied selection in their breeding programmes (Hedhammar 2020). Leighton 1997 reported that the percentage of German Shepherds (N = 2,037) and Labrador Retrievers (N = 1,821) with hip dysplasia decreased over five generations of selection. Therefore, breed clubs are encouraged to monitor the occurrence of genetic disorders and develop breeding recommendations in order to reduce their impact on canine welfare.

Animal behaviour is influenced by both genetics and environmental influences. However, the capacity at which both of these factors shape an individual's personality is heavily debated (Levitt 2013). Humans have always chosen animals that express behavioural traits beneficial to them (van Rooy et al. 2019). For this reason, there is a history of breeding dogs for specific traits as working dogs to assist humans (Adams 2008). Appropriate behaviour is an essential feature of all assistance dogs. The ability to alter behavioural characteristics in future litters would be of significant value to all assistance dog organisations.

There are many studies that discuss the effect of genetic factors on the suitability of dogs as working dogs (Bray et al. 2021b; van der Waaij et al. 2008; M. Goddard et al. 1983). Calculating heritability for traits through test batteries can be useful when selecting various kinds of working dogs (Wilsson et al. 1997b). One of the earliest studies looking at the genetic factors affecting the suitability of dogs as guide dogs recorded a heritability success of 0.44 which was significant enough to anticipate progress from selection, in particular against fear (M. E. Goddard et al. 1982). The heritability of seven behavioural traits in German Shepherds was studied. It showed that there is a possibility to direct behavioural characteristics in a more favourable direction by choosing possible breeders per their breeding value which would then increase the genetic gain in regard to desired behavioural traits (Meyer et al. 2012). Behaviours are a product of the interplay of both genetic background and the environment, therefore, the genetic background of the dog is just an indicator of the traits that may be observed in the mature dog (K. L. Overall et al. 2014).

Many organisations around the world use breeding programmes and some claim significant success (K. Evans et al. 2013). There is evidence that purpose-breeding and EBVs can be useful tools in genetically improving their stock. The Swedish Dog Training Centre (SDTC) proved that German Shepherds (N = 1,002) and Labrador Retrievers (N = 467) bred by them had higher test battery results (index values) than privately bred German Shepherds (N = 308) and Labrador Retrievers (N = 330). The SDTC dogs scored higher for many of the characteristics assessed, although they were outscored for one, sharpness (Wilsson et al. 1997b). Guide Dogs UK use EBVs to improve the accuracy of selection decisions and relies on phenotypic information and a DNA test where available when selecting breeding stock (K. Evans et al. 2013). The Japan Guide Dog Association uses its own breeding programme with the aim of ensuring the consistent production of dogs with specific traits. Approximately 30-40% of dogs successfully become guide dogs (Mizukoshi et al. 2008). Although purpose-bred assistance dogs are believed to have a higher rate of success (M. E. Goddard et al. 1982), some organisations rely on the general population to obtain a portion or all of their dogs.

There is a large amount of variation in the success rates of both purpose and non-purpose-bred dogs in the literature. A study by the Belgium Assistance Dog Association showed that the success rate of dogs originating from the general population is 60% with three out of five dogs graduating (Bogaerts et al. 2019). A study by J. A. Serpell et al. 2016 reported a success rate of 30-40% across five assistance dog organisations in the USA. Bray et al. 2019 reported a success rate of 43% in their purpose-bred dogs over 13 years in an American organisation that produces various types of assistance dogs. The success rate for

purpose-bred guide dogs is similar in Japan to those found by (Mizukoshi et al. 2008). An early study by M. E. Goddard et al. 1982 analysed data from 1,000 dogs that were both purpose-bred and non-purpose-bred. Over a period of 12 years, a guide dog organisation in Australia enrolled more purpose-bred puppies into their programmes and the success rates rose from 24% to 40% indicating the value of breeding programmes.

### 2.4.3 Training Considerations

After weaning, usually, at the age of eight weeks, puppies are typically given to a puppy raiser (Ennik et al. 2006). Puppy raisers are volunteers who foster purpose-bred prospective working dogs until they are mature enough for early and advanced training (D. Mai et al. 2020). Most recognised assistance dog organisations rely on puppy raisers to rear their puppies. A puppy's experience during their time with their puppy raisers is critical for their behavioural development (D. Mai et al. 2020). A puppy raiser's role is to teach the puppies basic obedience and socialisation skills (Ennik et al. 2006). This involves introducing novel stimuli to decrease the chance of fear-based responses later in their lives. Puppy raisers are given recommendations and support from their organisations on how to complete their role efficiently (L. Batt et al. 2008a).

With appropriate training recommended by the host organisation, the development of unsuitable behaviours can be avoided. Experienced puppy raisers often rear puppies that are less aggressive and less fearful when approached by strangers or unfamiliar dogs, while puppy raisers that are less experienced often rear puppies that have body sensitivity to touch which is an avoidance behaviour that is an undesirable trait for assistance dogs (J. A. Serpell et al. 2016).

Dogs are typically returned to their organisation between 14 and 18 months old (Ennik et al. 2006); however, some dogs can be as young as 12 months (Mizukoshi et al. 2008). Those that are not candidates for the breeding programme are neutered. Organisations have different methods for assessing their dogs' behaviour when they return from their puppy raisers. It is then decided if the dog is suitable for the training programme, or breeding stock or if it will be rejected (Ennik et al. 2006).

Some organisations allow their dogs to be "passed back", meaning that if a dog has not mastered some tasks or has some behavioural issues that must be addressed, they return to an earlier stage of training instead of being withdrawn from the programme. This lengthens the normal training period (Ennik et al. 2006), therefore training and graduation age can vary.

### 2.4.4 Discussion on Selection and Training Considerations

It is clear that selection and training considerations including breed, genetics, and training programmes influence the probability of success of trainee dogs. These factors should be considered when analysing the results of suitability or performance assessment methods.

In terms of breeds, Golden and Labrador Retrievers are established as suitable for assistance work. Geoffrey Caron-Lormier et al. 2016 validates this statement; however, the study was limited to 4 breeds. Organisations such as Canine Companions for Independence (CCI) (MacLean et al. 2018; Bray et al. 2019), Guide Dogs NSW/ACT (L. Batt et al. 2008b; L. M. Tomkins et al. 2011) primarily use Golden and Labrador Retrievers. The Swedish Dog Training Centre (SDTC) and The Seeing Eye (TSE) are inclined to include German Shepherds in their programmes with Golden and Labrador Retrievers (Wilsson et al. 1997a; Bray et al. 2017a). However, as the research on the suitability of these popular breeds increases, future research should focus on the suitability of Labradoodles, Goldendoodles, and other breeds that are becoming increasingly popular as assistance dogs.

Organisations such as CCI (MacLean et al. 2018; Bray et al. 2019), SDTC (Wilsson et al. 1997a), TSE (Bray et al. 2017a) and Guide Dogs NSW/ACT (L. Batt et al. 2008b; L. M. Tomkins et al. 2011) source dogs through their own breeding programmes. Although this indicates the success of breeding programmes in controlling health and behavioural issues as supported by scientific evidence (K. Evans et al. 2013), it would be unjust to ignore a large amount of variation in the success rates of both purpose and non-purpose-bred dogs. A possible explanation for these varying success rates across organisations may be that their screening methods differ. Some organisations could be stricter than others (Bogaerts et al. 2019). If it was possible to match organisations by their screening methods, and then make comparisons between purpose and non-purpose-bred dogs the results may be more accurate.

Even though assistance dog training programmes in different organisations share a similar structure, success rates vary significantly. They can be affected by multiple factors, including the presence or absence of a breeding programme; the available supply of puppies; and the demand for fully-trained canines. The type of behaviour assessment performed when the dogs return from their puppy raisers (Ennik et al. 2006), and the strictness of the evaluation (Bogaerts et al. 2019) may also cause success rates to vary. Hence, it is recommended that organisations be matched by their screening methods before making comparisons on success rates. Puppy raising offers a great opportunity to improve the effectiveness of training programmes due to its long-lasting impact on the dog's development (D. Mai et al. 2020). Future research should

investigate the stages of dog development, this would allow alterations to be made during puppy raising to address potential behavioural issues.

## 2.5 Behavioural Assessment Methods

In the most complete published review on temperament and personality research in dogs to date, Amanda C. Jones et al. 2005 identified four main methods along with their popularity amongst the studies examined. In particular, the methods were named Test Batteries (33%), Rating of Individual Dogs (18%), Expert Ratings of Breed Prototypes (18%), Observational Tests (16%) and the Others (15%) consisted of combinations of the four described assessment methods. More recently, Bray et al. 2021a systematically summarised all empirical studies that investigated the association between behaviour and work-related outcome in trainee dogs. The studies were categorised according to the behavioural measurement methods employed.

- **Test Batteries** (Amanda C. Jones et al. 2005) or **Behavioural Assessment** (Bray et al. 2021a): take place in a controlled environment where the presentation of stimuli and the evaluation of the dog's reaction to the stimuli takes place in a systematic way. They are generally composed of two components: the sub-tests and a coding system used for recording and scoring the subject's reactions.
- **Ratings of Individual Dogs** (Amanda C. Jones et al. 2005) or **Ratings** (Bray et al. 2021a): concern the individual behaviours and histories of a dog, they usually rely on the owner's subjective perception to complete a questionnaire.
- **Expert Ratings of Breed Prototypes** (Amanda C. Jones et al. 2005): focus on ranking breeds for different behavioural traits based on questionnaires filled out by dog experts.
- **Observational Tests** (Amanda C. Jones et al. 2005): take place in carefully selected, but not controlled, environments where naturally occurring stimuli are present, allowing inferences about a dog's temperament and behaviour pattern. Similar to Test Batteries, these tests are composed of the test itself and the system for scoring specific temperament traits based on an observed dog's overall performance.

There is a lack of standardisation in terminology related to dog behavioural tests (Rayment et al. 2015). Despite the similarities in the behavioural patterns assessed and the methodologies used, terminology varies depending on the authors. For example, test batteries have also been called behaviour tests and ratings have also been referred to as questionnaires. Brady et al. 2018 suggests the terms "character" tests for

assessing specific behaviour, “personality” tests for those assessing traits in agreement with human personality, and “temperament” tests for personality tests aimed at evaluating emotion rather than cognition. Canine temperament test batteries are often used to identify personality traits in order to assess a dog’s suitability for work. Researchers have mainly focused on stress response monitoring to develop personality models for non-human animals to describe coping-related behavioural patterns (Amanda C. Jones et al. 2005).

This review of the literature indicates canine activity monitors as a potential tool for assessing working dog behaviour for the first time. It provides a systematic comparison between the methods used in a number of renowned test batteries and ratings along with their applications and also discusses the use of various observational tests as a method of assessing behaviour. The following subsections consider four methods for assessing canine behaviour in the context of assessing suitability for work, namely test batteries, ratings, observational tests, and canine activity monitors. Each subsection contains the grouping definition and a review of the published literature given the inclusion criteria.

### 2.5.1 Test Batteries

Test batteries, also known as behaviour tests (Amanda C. Jones et al. 2005) or behavioural assessments (Bray et al. 2021a), take place in a controlled environment where stimuli are systematically presented to an animal according to a test protocol and their behaviour is assessed using a scoring system. The behaviour exhibited by the dog during the test is either recorded at the time or subsequently determined using video footage (Harvey et al. 2016b). The animal’s behaviour scores are statistically compared with another individual in the same situation with the aim of classifying the tested animal considering the goal of the test (Amanda C. Jones et al. 2005). They are considered to be less subjective than the rating of individual dogs and observational tests.

Test batteries have been developed and applied in various ways to assess canine behaviour for a range of purposes. Notably, behavioural issues are one of the most common reasons for trainee dogs to be withdrawn from their programmes (J. A. Serpell et al. 2001). Many of the dogs bred and trained are withdrawn due to the behavioural demands of their role. This decision often does not happen until the final weeks of their training programme. The resources required for breeding, raising, and training these dogs means that predicting the success of a dog in their training programme is crucial (Bray et al. 2019). Behaviour assessments that are capable of identifying unsuitable dogs for the role that they are being trained for are essential for assistance dog organisations due to the high percentage of fail rates caused by

behavioural reasons (Duffy et al. 2012). Therefore, several organisations have developed or adapted a number of test batteries as a performance assessment tool with the ultimate goal of predicting training outcomes (Amanda C. Jones et al. 2005; Bray et al. 2021a).

It is important to evaluate the quality of the procedures used to predict the success of working animals, with regard to their performance in their working role. There is a lack of information relating to the reliability and validity of test batteries (Brady et al. 2018). Like other techniques, there are some downsides to test batteries as they rely on the assumption that dogs would react in the same way independent of the context; in other words, they assume that the dog's reaction during the test battery is similar to that which they would display in other environments. For this reason, validation of these test batteries is necessary, but not always performed (Taylor et al. 2006). Valid and reliable test batteries are essential to organisations. Having an effective test battery could enable a more appropriate way of selecting dogs suitable for the role and choosing which dogs should be withdrawn (Brady et al. 2018).

Table 2.1 summarises 10 published relevant studies that have assessed the behaviour of candidate guide and/or assistance dogs using test batteries. It is ordered according to the different stages of development when the test batteries were conducted, starting from puppies, then juveniles and finally adults. It provides in-depth details of the test battery including the sub-tests performed and the variables scored. The variables that showed statistically significant differences between success and failure in the training programme are highlighted and discussed.

#### 2.5.1.1 Puppy Testing

Testing that occurs before the dog reaches 3 months is defined as puppy testing (Asher et al. 2013). Puppy testing can be highly beneficial to host organisations as they may not get the chance to assess the behaviour of these dogs until they return from their puppy raisers (J. A. Serpell et al. 2001). It has often been referred to as the holy grail of temperament testing due to its potential value of predicting the future behaviour of dogs from such a young age (Harvey et al. 2016b). Published literature on puppy testing reveals little consistency in the stimuli used, the age at testing, and the form of evaluation or validation (Asher et al. 2013).

The main disadvantage is that the test must be done while dogs are developmentally immature, therefore results obtained reflect the puppies' performance specific to the time of testing (J. A. Serpell et al. 2001). Between 6 and 8 weeks of age, puppies are motivated to approach strangers in contrast with the greater wariness before and after this age (K. M. Evans et al. 2015). Moreover, puppies have not fully developed the

fear imprinting response between the ages of 6 and 7 weeks, which facilitates certain testing as they are not fearful of unknown people (Alberghina et al. 2020).

Even though early life learning plays a role in their behaviour as adults, conducting test batteries on puppies aged 7 weeks of age is too early to be a reliable prediction of the future behaviour of a dog (Alberghina et al. 2020). Only a few studies show associations between puppy test results and training outcomes of working dogs (Slabbert et al. 1999; Svobodová et al. 2008; Asher et al. 2013). The lack of success in this area could be caused by neural and behavioural changes during the juvenile period which stabilise at social maturity (K. Overall 2013).

Table 2.1 describes three puppy tests that were trialled by organisations in an attempt to improve selection for their respective programs (K. M. Evans et al. 2015); namely, Puppy Profiling Assessment (PPA) (Asher et al. 2013), a test battery created by (Wilsson et al. 1997a) and Dog Cognitive Development Battery (DCDB) (Bray et al. 2020) - which was designed based on an adult test battery the Dog Cognition Test Battery (DCTB) (MacLean et al. 2017).

Table 2.1: Test Batteries used for assessing apprentice assistance dogs.

| Title<br>Publication | Affiliation<br>Programme<br>Breed | N<br>Age<br>Target | Sub-test | Variables |
|---|---|---|---|---|
| | | Puppy Testing | | |
| Puppy Profiling Assessment (PPA)<br>Asher et al. 2013 | Guide Dogs for the Blind Association (GDBA)<br>Guide<br>LR; GR; LR X GR; GS; GR X FR; GR X GS; CR; FR | 465<br>6-8 Weeks<br>Effects of behaviour test measures on training outcome | (1) Following when called (2) Retrieve a toy (3) Gentle restraint (4) Noise (5) Stroking (6) A model moving animal (squirrel) (7) A tunnel (8) A ramp | (1) Following (2) Response to stimuli, Response to assessor* (3) Restraint (4) Noise test (5) Response to stimuli*, Response to assessor* (6) Response to stimulus, Response to assessor* (7) Tunnel (8) Ramp* |
| NA<br>Wilsson et al. 1998 | Swedish Dog Training Centre (SDTC)<br>Guide<br>GS | 630<br>8 Weeks and 15-20 Months<br>Relationship between behaviour measures using Wilsson et al. 1997a, adult behaviour and heritability of traits | Part I: Vocalisation when isolated. (1) Yelp (2) Shriek (3) Contact I (4) Fetch (5) Retrieve (6) Large ball (7) Tug of war. Part II: Arena test. (8) Contact II (9) Activity (10) Objects visited | (1) Yelp (2) Shriek (3) Contact I (4) Fetch (5) Retrieve (6) Large ball (7) Tug of war (8) Contact II (9) Activity (10) Objects visited |

Continued on next page

Table 2.1 - continued from previous page

| | | | | |
|---|---|---|---|---|
| Dog Cognitive Development Battery (DCDB) Bray et al. 2020 Bray et al. 2021c | Canine Companions for Independence (CCI) Assistance LR X GR; LR; GR | 168 8-10 Weeks Characterise the early development of dog cognition | Vision pretest: (1) Retrieval (2) Laterality. Hiding-finding warm-ups 1: (3) Human Interest (4) Cylinder (5) Unsolvable. Hiding-finding warm-ups 2: (6) Communicative marker (7) Arm pointing (8) Odour control (9) Novel object. Hiding-finding warm-ups 3: (10) Working memory (11) Visual discrimination (12) Auditory discrimination (13) Odour discrimination (14) Surprising event | (1) Average score, Tally (2) Bias strength, laterality index. (3) Average look time, Average interact time (4) Familiarisation score, Inhibitory Control: Score, Latency; Reversal Learning: Score, Latency, First side correct (5) Average time looking at human, Average time manipulating object. (6) Communicative marker (7) Arm pointing (8) Odour control trials. (9) Memory: 5s, 10s, 15s, 20s (10) Visual discrimination (11) Auditory discrimination (12) First choice, Final choice, Persistence, Time at the correct response, Time at incorrect response |

Continued on next page

Table 2.1 - continued from previous page

| Juvenile Testing | | | | |
|---|---|---|---|---|
| NA<br>Harvey et al.<br>2016b | NA<br>Guide<br>GR X LR, LR, LR X GR, L X GRx, GS X GRx | 93<br>5 and 8 Months<br>Effects of behavioural test measures on training outcome | (1) Meet a stranger (2) Obedience with PW (3) Obedience with a stranger (4) Raised path (5) Body check (6) Head ring (7) Tea towel (8) Food (9) Robin (10) Pigeons (11) Human distraction | All: Jumps, Whines, Scratches, Barks*, Shakes*, (1) Pull strength, Low posture* (2 & 3) Obedience: Sit/Wait/Down* response, Gaze proportion, Lip licks* (4) Path: Crossed (5) Body check: Score, Mouths, Licks (6) Head ring: Ear position, Tail height, Body posture (7) Tea towel: Attempts to remove*, Turns head*, Change from neutral*, Plays with* (8-10) Distractions: Pull strength*, Time oriented*, Approaches*, Avoids* (11) Human distraction: Pull strength, Jumps |
| NA<br>L. Batt et al.<br>2008b | Guide Dogs NSW/ACT<br>Guide<br>LR; GR | 43<br>6 and 14 Months<br>Effects of temperament, lateralization, cortisol on training outcome | (1) Social contact (2) Passive (3) Chase (4) Noise (5) Dog distraction (6) Sudden appearance | (1) Latency to approach tester (2) Latency to Drop*/Rest*/Sit (3) Latency to catch, Duration of hold (4) Latency to recover from noise 1/2/3/4/5 tests (average) (5) Jumping up*, Wag, Bark, Occurrence of Whining/Pulling |

Continued on next page

Table 2.1 - continued from previous page

| | | | | |
|---|---|---|---|---|
| | | Adult Testing | | |
| NA<br>Wilsson et al.<br>1997a | Swedish Dog Training Centre (SDTC)<br>Guide<br>LR; GS | 2107<br>15-20 Months<br>Effects of behaviour test results on future capacity as service dogs | (1) Approachability and tendency to compete for objects (2) Startling test I (3) Startling test II (4) Reaction to a loud noise (5) Reaction to a successively approaching threat (6) Attack on the handler (7) Reaction to gunfire | (A)* Courage (B) Sharpness (C) Defence drive (D) Prey drive (E)* Nerve stability (F)* Reaction to gunfire (G) Temperament or energy level (H)* Hardness (I)* Ability to cooperate (J)* Affability |
| NA<br>Bray et al.<br>2017a | The Seeing Eye (TSE)<br>Guide<br>GS, LR, GR | 98<br>14-17 Months<br>Effects of maternal investment, temperament and cognition measures on training outcome | (1) Isolation (2) Distraction (3) Sustained attention (4) Memory problem solving (5) Multistep problem solving (6) Cylinder (7) Detour problem solving (8) Greeting (9) Ball play (10) Novel object (11) Umbrella-opening | (1) Anxious, Active (2) Distraction (3) Sustained attention (4) Superior performance, Accuracy (5) Poor performance* (6) Cylinder (7) Poor performance (8) Greeting (9) Ball play (10) Wary, Quiet* (11) Reactivity*, Recovery |
| NA<br>L. M. Tomkins et al.<br>2011 | Guide Dogs NSW/ACT<br>Guide<br>LR, GR, LR X GR | 113<br>13-17 Months<br>Effects of kennel behaviour and temperament on training outcome | (1) Passive and Noise Test (2) Sudden Appearance Test (3) Dog Distraction Test | (1) Latency to sit*, Drop, Rest (2) Barking, Jumping, Shivering, Cowering, Raising of hackles (3) Occurrence of pulling, Panting*, Number of jumps, Tail wags, Barks, Whines, Forepaw lifts, Licks*, Avoidance steps taken away from the unfamiliar dog, Growls, Raising of hackles |

Continued on next page

Table 2.1 - continued from previous page

| | | | | |
|---|---|---|---|---|
| In For Training (IFT) Bray et al. 2019 | Canine Companions for Independence (CCI) Assistance LR, GR, LR X GR | 5967 18 Months Predictive power of behaviour test and questionnaire measures on training outcome | (1) Physical exam (2) Looming object (3) Sudden loud noise (4) 'Prey' object (5) Unfamiliar dog (6) Threatening stranger | (1) Ease of handling*, General demeanour, Body tension*, Vocalisation, Tail position (2) Initial reaction*, Recovery, Barks or growls, Increase in activity, Second walk by (3) Initial reaction*, Recovery, Conclusion*, Barks or growls (4) Initial reaction*, Recovery, Conclusion, Barks or growls (5) Initial reaction*, Recall*, Hackles, Barks or growls, Tail position (6) Initial reaction, Recovery, Hackles, Barks or growls, Increase in activity |

Continued on next page

Table 2.1 - continued from previous page

| | | | | |
|---|---|---|---|---|
| Dog Cognition Test Battery (DCTB) MacLean et al. 2017; MacLean et al. 2018 | Canine Companions for Independence (CCI) Assistance LR; GR; LR X GR. | 164 24 Months Effects of cognitive measures on training outcome | (1) Affect discrimination (2) Arm pointing (3) Causal reasoning (4) Contagious yawning (5) Cylinder (6) Detour navigation (7) Gaze direction (8) Hiding-finding (9) Inferential reasoning (10) Laterality (first step) (11) Laterality (object manipulation) (12) Marker cue (13) Memory-distraction (14) Odour control trial (15) Odour discrimination (16) Perspective- taking (17) Reaching (18) Retrieval (19) Reward preference (20) Rotation (21) Sensory bias (22) Social referencing (23) Spatial preservation (24) Spatial transpositions (25) Transparent obstacle (26) Unsolvable task (27) Visual discrimination (28) Working memory | (1) Affect discrimination (2) Arm pointing* (3) Causal reasoning: Visual, Auditory* (4) Contagious yawning (5) Cylinder* (6) Detour navigation (7) Gaze direction (8) Hiding-finding (9) Inferential reasoning* (10) Laterality (first step)* (11) Laterality (object manipulation)* (12) Marker cue (13) Memory- distraction (14) Odour control trials (15) Odour discrimination* (16) Perspective- taking: Obey command, Steal food (17) Reaching (18) Retrieval (19) Reward preference* (20) Rotation (21) Sensory bias (22) Social referencing* (23) Spatial preservation (24) Spatial transpositions* (25) Transparent obstacle (26) Unsolvable task: Look at experimenter*, Manipulate container (27) Visual discrimination (28) Working memory |

Dog Breeds: Labrador Retriever (LR); Golden Retriever (GR); Labrador Retriever x Golden Retriever (LR x GR); Golden Retriever x Labrador Retriever (GR x LR); Labrador Retriever x Golden Retriever x (LR x GRx); Golden Retriever x Flat Coated Retriever (GR x FR); Curly Coated Retriever (CR); German Shepherd (GS); German Shepherd x Golden Retriever (GS x GRx). Dam/Sire crosses and generations of crosses were not included in this table, only the breeds of dogs used.
Variables with an asterisk(*) above them indicate that the variable was significant in determining success or failure.

The test battery designed by Wilsson et al. 1998 and DCDB designed by Bray et al. 2020 did not provide details of any statistically significant variables that could possibly identify the puppies' success or failure in their programme as the puppies' training outcome was not discussed. There was no reliability or validity studies completed in Wilsson et al. 1998 and only inter-rater reliability was reported for Bray et al. 2020. The Puppy Profiling Assessment (PPA) was developed using similar stimuli to the Puppy Aptitude Test (PAT), but with a new and possibly more effective scoring protocol. The new scoring protocol measures responsiveness, confidence reactions to the human assessor, and reactions to environmental stimuli (Asher et al. 2013). The PPA had a number of statistically significant variables and also had the reliability of the test assessed (Asher et al. 2013).

The Campbell Test and Puppy Aptitude Test (PAT) (designed based on the Campbell Test) were once popular in the literature, but are not included in Table 2.1 due to lack of information on their original creation. Although modified versions of the Campbell Test and PAT are used in current studies (Alberghina et al. 2020), the original test assumed that dominance is a suitable trait which is widely considered to be false (Asher et al. 2013).

An evaluation of the Campbell Test showed no correlation between the test results recorded at 7 and 16 weeks of age, suggesting that it was not an effective test for predicting adult behaviour (Asher et al. 2013). A modified version of PAT measuring retrieving, sociability, and cooperation with people has showed some correlations with scores in an adult test battery (Asher et al. 2013).

#### 2.5.1.2 Juvenile Testing

The juvenile testing period occurs when the dogs are approximately 6 months of age and ends when the dogs reach sexual maturity (L. S. Batt et al. 2009). The juvenile period is probably the least well studied despite its potential of affecting adult behaviour (J. A. Serpell et al. 2016). However, there are a few studies which indicate that the behaviour of juvenile dogs can be partly predictive of their behaviour as an adult (Harvey et al. 2016b; Sforzini et al. 2009). Behavioural traits become more stable during the juvenile period (L. Batt et al. 2008b). Therefore, conducting behavioural testing methods later in the juvenile period could improve the tests predictive value. Because of the many developmental changes that puppies experience during their first 12 weeks, assessing juvenile or young adult dogs could be an effective method of predicting the success of the dog in the training programme (Harvey et al. 2016b).

Table 2.1 presents two juvenile dog test batteries created in an attempt to assess prospective guide dogs' suitability for work before starting

formal training.

Harvey et al. 2016b designed and evaluated a test battery consisting of 11 stimuli to assess the dog's reaction to meeting a stranger, obedience commands, body sensitivity, scavenging, and animal and human distractions. Trainee guide dogs participating in this longitudinal study took the test at 5 months (5M) and 8 months (8M) of age. Statistically significant associations between training outcome and 7 behavioural measures for 5M and 5 measures for 8M were found. Only one variable was significantly associated with training outcomes at both testing sessions. The time oriented towards food was associated with withdrawal from the programme. Dogs who spent a longer amount of time-oriented towards food had an increased chance of being withdrawn. The authors demonstrate the validity of this assessment method for predicting dogs' training outcomes with an accuracy of 79.7% and 87.3% for the 5M and 8M models, respectively. An acceptable intra-rater reliability score is presented while the inter-rater reliability could not be discussed as the tests were scored by a single rater.

In the test battery created by L. Batt et al. 2008b the three variables were proved to be statistically significant contributors to guide dog success; namely, a shorter latency to drop, a greater latency to rest, and an absence of jumping. The training outcome used to validate the test results was the dog's performance at the end of their training programme.

In conclusion, the test battery designed by Harvey et al. 2016b provided a number of variables indicative of success being capable of identifying dogs likely to succeed. On the other hand, the test battery presented in L. Batt et al. 2008b provided variables predictive of overall failure being more suitable for recognising dogs like probable to fail.

#### 2.5.1.3 Adult Testing

Although there are different opinions surrounding the onset of adulthood in dogs (Case 2005; Harvey et al. 2016b), dogs are returned to their host organisations when they are over-aged 1 year old for advanced training (Ennik et al. 2006). The accuracy of the predictions made based on behavioural assessments increases with age (L. Batt et al. 2008b). Performing a test battery when the dog is an adult rather than a juvenile is more successful in predicting the future behaviour of that dog. L. Batt et al. 2008b proposes 14 months to be the optimal age for behavioural assessment; indeed, most trainee dogs would have reached this age and be available for testing which is highly beneficial to the host organisations.

When it comes to assessing adult behaviour, assistance dog organisations carry out their test batteries. However, some tests are not

designed in conjunction with an animal behaviour specialist and they are not presented in the scientific literature. Often those that are presented in the literature do not report the quality of the test in terms of reliability and validity. Many papers do not describe the full details of the test procedure and how they score the dog's behaviour making it difficult to replicate, thus hindering scientific progress in this field (Taylor et al. 2006). However, some of the most popular adult test batteries in the scientific literature were designed by organisations and are reliable (Svartberg et al. 2002; Asher et al. 2013; Bray et al. 2019).

Five adult test batteries are discussed in Table 2.1; namely, Dog Cognition Test Battery (DCTB)(MacLean et al. 2017; MacLean et al. 2018), In-For-Training (IFT) (Duffy et al. 2008; Bray et al. 2019) and the three test batteries created by Wilsson et al. 1997b (I), Bray et al. 2017a, and L. M. Tomkins et al. 2011 respectively.

The Dog Cognition Test Battery (DCTB) designed by MacLean et al. 2017; MacLean et al. 2018 aimed to measure diverse cognition processes when assessing links between cognition and working dog performance. Additionally, it was also used to identify the specific links between these aspects of cognition and success in different working dog roles. Out of the 28 variables scored in this test battery, 11 were considered statistically significant. There are several sub-tests in this test battery which can make it unappealing for many researchers; however, reliability, inter-rater reliability and predictive validity have all been assessed. Predictions made using these statistically significant variables were much more accurate for the subset of dogs predicted to have the highest probability of success. Therefore, predictions were calculated separately for dogs with the highest predicted probability of success. This study found that it may be difficult to produce accurate predictions for all candidates; however, the measures and models created in this study may be valuable for identifying the subset of dogs with the most potential (MacLean et al. 2018).

The In-For-Training (IFT) is a standardised test battery used by several organisations when the dogs return from their puppy raisers to begin their professional training, which represents the biggest investment of resources in their career. The IFT is similar to test batteries used by various working dog groups in Sweden (Wilsson et al. 1997a) and the UK (Harvey et al. 2016b) and are scored by a small group of experienced trainers. It was first mentioned in a paper by Duffy et al. 2008 where they discuss its use as part of a project and relationship with C-BARQ questionnaire; however, no details on the ethogram scoring system were provided. The IFT was later discussed by Bray et al. 2019. Bray et al. 2019 investigated its predictive validity and suggested that outcome predictions based on the IFT may help streamline and expedite decisions regarding which dogs to retain for subsequent professional training or breeding purposes. Its capability of producing accurate

predictions about the dogs least likely to succeed and also identifying the dogs most likely to graduate their programme provides the researcher and trainers with a vast amount of information about their candidates, which will support them in making key decisions about the future career of the dogs.

The test battery reported by Wilsson et al. 1997a was conducted on both police and guide dog candidates. For the purpose of this study, only the results significant to Labrador Retrievers are discussed as according to the study, these are more likely to become guide dogs than German Shepherds. Four of the ten variables scored in this test battery were considered statistically significant and the test battery reliability was studied. Labrador retrievers were also found to be more cooperative and more affable. This study shows that this and other subjectively evaluated complex behavioural parameters can be evaluated with a high degree of reliability. However, the reliability of these results is inversely proportional to the number of people completing the observations. The authors believe the reliability of the test could be improved by repeating testing procedures and independent judgements by two or more people.

The test battery described in Bray et al. 2017a consisted of only 11 sub-tests and had 11 variables to score them. Three of these variables were associated with the dog's outcome in their training programme. This test battery focuses on problem-solving and temperament tasks. The results from this study indicate that forcing tasks into categories due to association with problem-solving or temperament leads to a poor fit. It also leads to factors that do not correlate with other important variables or predict outcomes.

L. M. Tomkins et al. 2011 created a test battery that differed from its predecessors as it measured the occurrence and the frequency of behaviours. It only has three subtests and a total of 12 variables to score them. Three variables were associated with the increased probability of the dog's success in the programme. This study did not provide details of any reliability or validity studies. The statistically significant variables were: the presence of panting, the presence of licking and the latency to sit during the third noise test. To be more specific, measures that increased the probability of success were prompt sitting in the third noise test and the absence of panting and licking during the dog distraction test. This test battery could be beneficial to many organisations as it is a short test. The authors believe that, in an experimental scenario, the removal of dogs that did not sit and did not pant and lick would increase success rates in dog training organisations.

There are other types of test batteries that use various methods to predict guide dog success with a more detailed description of these available in Bray et al. 2021a. For example, one study mentioned in

Table 1 by Bray et al. 2021a performed a series of tests to assess the three categories of laterality, motor, sensory and structural by L. M. Tomkins et al. 2012. Motor lateralisation was assessed using a Kong test to observe paw preference. Sensory lateralisation was assessed using a Sensory Jump Test (SJT) described in L. Tomkins et al. 2010 to observe visual biases. Structural lateralisation was assessed by classifying hair whorl characteristics in the dogs. The results from this study identified many significant predictors of success and supported the hypothesis that all three categories of laterality could be used to predict the dog's success in their training programme (L. M. Tomkins et al. 2012). A similar study was carried out on family pets using the Kong test to compare lateral bias (Wells et al. 2016). A more recent study by Wells 2021 discusses the various methods used for assessing preferential paw use in both dogs and cats. These studies highlight the variability in test batteries used in literature and described in Table 2.1.

### 2.5.2 Observational Tests

Observational tests are popular in scientific literature. The aim of observational testing as described by Amanda C. Jones et al. 2005 is to "assess and describe relatively broad traits discernible in naturalistic environment". They are similar to test batteries as they have two major sections: the test and the scoring system. They differ from test batteries in that they are conducted in carefully selected but not controlled environments (Amanda C. Jones et al. 2005). They involve experienced trainers observing a dog's behaviour under natural or semi-natural conditions. They have been conducted on apprentice guide dogs in order to further define temperament categories in terms of observable behaviour (Murphy 1998) and as potential predictors of training outcome (Mizukoshi et al. 2008; L. M. Tomkins et al. 2011).

An observational test was conducted by Murphy 1998 at the Royal Guide Dogs Association of Australia (RGDAA) with the aim of developing descriptions of the temperament categories used by the trainers. Dogs were video recorded while walking along a defined route by their trainer for five consecutive days. Trainers described their behaviour using 20 temperament categories and scored 107 behaviour items based on the video recordings. The goal was to understand the link between observable behaviour items and the 20 temperament categories. This study suggests that it is beneficial to have sets of behavioural characteristics to assist in the recognition of temperament categories as a single behaviour is only a component of a whole behaviour pattern. However, several of the 20 temperament categories were rarely seen or never occurred. Moreover, different temperament categories also share the same behavioural elements such as excitability, suspicion and dog distraction shared the element rigid tail. The association between

behaviour, temperament and training outcome was not investigated.

Another observational test was conducted by Mizukoshi et al. 2008 at the Japan Guide Dog Association. It was designed with the aim of assessing task performance, stress, excitement and concentration. A sample of eight apprentice guide dogs was video recorded every two weeks for the first 12 weeks of formal training while completing an observation course in a residential area. Although only one dog became a guide, these results could be used as a stepping stone for future research. The four factors monitored demonstrate that the successful dog took less time to complete the course, showed fewer stress reactions as training progressed, less time distracted by obstacles, and more horizontal tail position.

L. M. Tomkins et al. 2011 conducted an observational test in the Guide Dogs NSW/ACT Training Centre to explore behaviours shown during the time spent in kennels. The authors hypothesised that behaviour measures observed between 5 pm to 8 am could be used for the early identification of successful dogs. The behaviour of five dogs in five different groups (N = 25) was assessed for one minute every ten minutes during the 15h of surveillance on days 1,2,3,7 and 14 after admission to kennels. A range of behaviours was observed; however, only one was significantly correlated with training outcome, specifically, the longer the dog spent resting the higher the odds of success.

### 2.5.3 Ratings

Ratings are essentially questionnaires consisting of either context or adjective-based items that are assessed through behavioural scales or behavioural codings (Barnard et al. 2016). They are widely used for evaluating behavioural traits in potential assistance dogs as a tool for assessing their suitability for the training programme (J. A. Serpell et al. 2001).

There are a variety of ratings to choose from, each specifically designed to assess certain aspects of dog personality and behaviour (Rayment et al. 2016; Bray et al. 2021a). Ratings are used to identify key personality traits in pet dogs such as impulsivity (Hannah F Wright et al. 2011) and aggressiveness (Mirkó et al. 2012). These rating methods could be beneficial to assistance dog organisations. They offer an alternative or act as a complement to the standardised test battery (Wiener et al. 2016). In fact, a number of studies comparing results from test batteries and questionnaires have been carried out in detection (Lazarowski et al. 2021), military (Foyer et al. 2014), shelter (Clay et al. 2020), assistance (Dollion et al. 2019), and pet dogs (Barnard et al. 2016; Siracusa et al. 2010).

The availability of ratings through the internet has caused them to

become a popular method of implementing citizen science (Wiener et al. 2016). Many organisations rely on the experience and judgement of people such as puppy raisers who have lived with the dogs or trainers who have worked with the dogs to complete the questionnaire. Raters consider the dogs' behaviour in situations and environments that are natural to them and reflect their typical response over time (Duffy et al. 2012).

Most ratings ask questions or provide statements for an assessment of the dog's behaviour in specific situations, rather than assigning personality adjectives. In order to analyse behaviour patterns, rating items must be refined to a small number of workable constructs. Validity and reliability of the results obtained from ratings are important and can be determined with different statistical tools (Wiener et al. 2016).

Table 2.2 presents seven studies that have applied and/or developed ratings to assess the behaviour of assistance dogs. They were categorised according to their main goal, including canine personality assessment and suitability for assistance work. Each of these ratings was used on various dog breeds; however, most importantly were used on working dog breeds. Table 2.2 details the number of dogs involved in the study, the questionnaire items, the rating method, and the scales used for scoring.

#### 2.5.3.1 Canine Personality Assessment

The Canine Behavioral Assessment and Research Questionnaire (C-BARQ) is the most commonly used rating (Wiener et al. 2016). It is standardised and validated (J. A. Serpell et al. 2016). It was originally developed to identify troublesome behaviours in dogs within a clinical setting. It is currently used by many organisations worldwide (Duffy et al. 2012) for a variety of purposes such as assessing canine personality traits (Wiener et al. 2016), gathering information from dogs in shelters (Clay et al. 2020) and differentiating breed tendencies in phenotypic studies (van Rooy et al. 2019). A significant relationship was found between C-BARQ scores based on the dog's behaviour between 6 and 12 months of age and their success in guide dog training (Duffy et al. 2012). It was also used in the broad-scale assessment of behavioural traits of dogs from different countries, moreover, it was deployed as a tool to validate test batteries (Rayment et al. 2016; Foyer et al. 2014; Lazarowski et al. 2021).

Table 2.2: Rating systems used to assess canine personality or suitability for assistance work.

| Questionnaire title<br>Publication | N<br>Scorer | Items<br>Rating | Scales |
|---|---|---|---|
| **Canine Personality Assessment** | | | |
| Canine Behavioral Assessment and Research Questionnaire (C-BARQ)<br>Hsu et al. 2003; J. Serpell 2021 | ∼7500<br>Pet dog owners | 101-item context-specific<br>0 to 4 | Training and Obedience, Aggression, Fear and Anxiety, Separation-related behaviour, Excitability, Attachment and attention-seeking, Miscellaneous |
| Monash Canine Personality Questionnaire - Revised (MCPQ-R)<br>Ley et al. 2009a; Ley et al. 2009b; Ley et al. 2008 | 450<br>Pet dog owners | 26-item adjective-based<br>1 to 6 | Extraversion, Motivation, Training Focus, Amicability, Neuroticism |
| Dog Personality Questionnaire (DPQ)<br>Amanda Claire Jones 2008 | >5,000<br>Volunteers | 45 or 75-item context-specific<br>1 to 7 | Fearfulness, Aggression towards People, Activity/Excitability, Responsiveness to training, Aggression towards Animals |
| **Suitability for Assistance Work** | | | |
| NA<br>Arata et al. 2010 | 144<br>Guide dog trainers | 22-item adjective-based<br>1 to 5 | Three factors: Distraction, Sensitivity, Docility. Six items: Initiative, Aggression, Animal interest, Food interest, Scent interest and Suspicion |

Continued on next page.

Table 2.2 - continued from previous page.

| Questionnaire title<br>Publication | N<br>Scorer | Items<br>Rating | Scales |
|---|---|---|---|
| Revised Puppy Walker Questionnaire (r-PWQ)<br>Harvey et al. 2016a; Hunt et al. 2020 | 276<br>Puppy raisers | 39-item context-specific 10cm line | Separation related behaviour, Excitability, Attachment and attention seeking, Energy level, General anxiety, Animal chase, Body sensitivity, Distractibility and Trainability. |
| Puppy Training Supervisor Questionnaire (PTSQ)<br>Harvey et al. 2017 | 1,401<br>Guide dog trainers | 39-item context-specific 10cm line | Trainability, General Anxiety, Adaptability, Excitability, Body sensitivity, Distractibility, Stair anxiety, Miscellaneous |
| Questionnaire for Foster Families (QFF)<br>Dollion et al. 2019 | 3,440<br>Puppy raisers | 25 and 37-item context-specific categorical | House Training, Activity/Behaviours, Attractions, Fears |

The preliminary versions of this rating were developed in the early 2000s by Hsu and Serpell. The original version of this questionnaire had 8 factors and 40 questions (J. A. Serpell et al. 2001). It then evolved to have 11 factors and 68 questions (Hsu et al. 2003). By 2005, it had stabilised at 14 factors and 100 context-specific questions (Duffy et al. 2012; J. Serpell et al. 2005). Raters grade either the frequency or severity of behaviour in certain situations on a 5-point scale (Wiener et al. 2016). It is advantageous for organisations to have the C-BARQ completed by puppy raisers as it produces an idea of the dog's behavioural profile before they begin advanced training (Bray et al. 2019). It provides details on the characteristics of the household, the puppy raiser's experience and the puppy's exposure to many environmental stressors and stimuli (J. A. Serpell et al. 2016).

The Monash Canine Personality Questionnaire (MCPQ) was originally created by Ley et al. 2008, then it was refined by Ley et al. 2009b and named Monash Canine Personality Questionnaire - Revised (MCPQ-R). It was developed using an adjective-based methodology similar to the ones used in the Big Five Model (BFM), a model of human personality. The MCPQ-R has five personality dimensions: Extraversion, Training focus, Amicability, Neuroticism, and Motivation (Ley et al. 2009b). The original version of this rating had 67 personality adjectives, it then went to 41 and has now stabilised at 26 (Wiener et al. 2016). The goal of this rating method was to create a "psychometrically sound instrument" that could be used to assess a dog's personality (Ley et al. 2009a). The reliability of this method was tested using inter-rater and test-retest reliability measures (Ley et al. 2009a). This rating was used on 82 dog breeds which were allocated into seven groups: toys, terriers, gundogs, hounds, working dogs, utility, and non-sporting. Differences in the five personality sub-scale scores across the groups were statistically significant. Working dogs and terriers scored significantly higher on the extroverted sub-scale compared to other groups, and working dogs along with gundogs also scored highly on the training focus sub-scale. Inter-rater and test-retest reliability analysis showed good agreement between different raters and temporal stability, respectively. Therefore, the MCPQ-R can be considered a valid and reliable method for assessing canine personality (Ley et al. 2009a).

The Dog Personality Questionnaire (DPQ) was created by Amanda Claire Jones 2008 to provide assistance dog programmes with a tool capable of quickly measuring canine personality. The DPQ is available as a 45-item or 75-item rating method. The dogs are rated on five scales with a score of 1-7. Each scale contains a number of sub-categories. For instance, Activity/Excitability is made up of Excitability, Playfulness, Active Engagement, and Companionability (Posluns et al. 2017). It has been extensively tested, resulting in acceptable levels of inter-rater reliability, test-retest reliability, and

predictive validity (Amanda Claire Jones 2008). The DPQ was used in several studies since its development (Wood 2020).

#### 2.5.3.2 Suitability for Assistance Work

The Puppy Training Supervisor Questionnaire (PTSQ) was created by Harvey et al. 2017 with the purpose of predicting the outcome of guide dogs in training as they could not change items in the C-BARQ. The PTSQ contains 39 items that are context-specific questions. Thirty-six of these items form nine targeted traits and three of these items are miscellaneous items. The majority of these items are behavioural descriptions; however, some are adjective items. The PTSQ contains seven scales; namely, Adaptability, Body sensitivity, Distractibility, Excitability, General anxiety, Trainability, and Stair anxiety. The dogs are scored on a 100 mm long visual analogue scale ranging from 'Never' and 'Almost Always' for behavioural description items and 'Really does not describe this dog' to 'Really describes this dog' for adjective items.

The PSTQ demonstrated temporal consistency, construct, concurrent and predictive validity; however, inter-rater and intra-rater reliability were not assessed in this study. The creation of a flag system to aid decision-making allowed the use of scores given to dogs as early as 5 months of age to identify those most likely to succeed and scores are given at 8 and 12 months of age to identify dogs at high risk of withdrawal. It is believed that this questionnaire is the first of its kind with the ability to predict the training outcome of individual dogs that are most and least suitable to the role of a guide dog (Harvey et al. 2017). A recent study was conducted by D. L. Mai et al. 2021 which used the PSTQ to assess specific puppy-raising practices that could predict desirable behaviour in trainee assistance dog puppies. It found that socialisation and training, which are two of the roles that puppy raisers take on, were strong predictors of the puppy's behaviour.

Another rating method was used in a study at the Japan Guide Dog Association (JGDA) by Arata et al. 2010. It was used at JGDA since 2002 and originates from the Guide Dogs for the Blind Association of Queensland. Trainers at JGDA completed the survey on 144 dogs that completed three months of their training. The dogs are assessed based on 22 adjective items scored on a 1-5 scale where five is the highest degree of an item. The aim of the study by Arata et al. 2010 was to examine the association between temperament traits and qualification, examine the reproducibility of results, and the impacts of temperament traits on the prediction of success. Factor analyses identified lower Distraction, Sensitivity, and higher Docility as important behavioural factors for guide dogs when determining the training outcome of that dog. They presented high internal consistency and could be assessed stably by the rating method. Additionally, the other six items - namely,

higher Initiative and lower Aggression, Animal interest, Food interest, Scent interest, and Suspicion - had signification associations with training outcomes after Bonferroni correction. The significant traits found in this study are similar to those found in M. Goddard et al. 1983 and J. A. Serpell et al. 2001. This suggests the results found in this study have validity (Arata et al. 2010).

The Puppy Walker Questionnaire (PWQ) was developed for puppy raisers. It originally consisted of 59 items which were sourced from published questionnaires such as (M. Goddard et al. 1983; J. A. Serpell et al. 2001; Hsu et al. 2003; Arata et al. 2010), puppy walker ratings from an unpublished study undertaken by the Guide Dogs UK or created to address certain behavioural traits (Harvey et al. 2016a). Out of the 59 items, 22 of them were sourced from the C-BARQ (Harvey et al. 2016a). The PWQ was later revised by Hunt et al. 2020 and is now referred to as the r-PWQ (refined Puppy Walker Questionnaire). The r-PWQ consists of 39 items in nine scales, which are Separation related behaviour, Excitability, Attachment and attention seeking, Energy level, General anxiety, Animal chase, Body sensitivity, Distractibility, and Trainability. These items are scored using a 0 to 100 scale where zero means “Never” and 100 means “Almost always”. Inter-rate reliability analysis resulted in acceptable values (Harvey et al. 2016a; Hunt et al. 2020).

The Questionnaire for Foster Families (QFF) was created by Dollion et al. 2019 and used in the Mira Foundation. It consists of 25 items for dogs aged 6 months and 37 items for those aged 12 months. The answers for these items are coded into categorical scores. The internal consistency of factors and associations between the scores of each dog for Fear/Reactivity was assessed. Results indicated that the Fear/Reactivity factor at 6 months of age was significantly correlated with both the Fear/Reactivity factors at 12 months of age. Regression analyses showed that Fear/Reactivity is predictive of dogs’ disqualification (Dollion et al. 2019).

As the most commonly used rating method, the Canine Behavioral Assessment and Research Questionnaire (C-BARQ) is suitable for assistance dogs (Wiener et al. 2016). This statement is supported by the fact that it is used in many organisations worldwide (Duffy et al. 2012). The C-BARQ was used in many studies. The Monash Canine Personality Questionnaire (MCPQ) is also an appropriate rating method for assistance dogs as it provides valid and sensible measures of dog personality (Ley et al. 2009b), and it iS reliable (Ley et al. 2009a).

There are many different ratings in literature with those in Table 2.2 only highlighting a few. It is important that researchers understand what they want the outcome of their study to be, and what traits they want to assess and choose a rating method based on that.

### 2.5.4 Mixed Assessment Methods

Some of the research available includes both a test battery and a rating in their methods. For the purpose of this review of the state of the art, these are called "Mixed Assessment Methods". Mixed Assessment Methods are useful in identifying the validity and reliability of the behavioural assessment method.

Svartberg 2005 compared the C-BARQ results (N = 697) completed by dog owners to the Dog Mentality Assessment (DMA), a test battery conducted on police dogs in the Swedish Dog Training Centre (SDTC). Wiener et al. 2016 showed that C-BARQ ratings had a statistically significant correlation with four behaviour traits from DMA; namely, playfulness, curiosity/fearlessness, sociability, and distance playfulness. At Canine Companions for Independence (CCI), combining the C-BARQ ratings completed by puppy raisers with the IFT scores (determined by a small group of trained evaluators) improved the predictor model's accuracy in the lowest 20% category; hence, trainee dogs not suitable for the assistance programme were more accurately identified (Bray et al. 2019).

The DPQ was compared with MCPQ-R and the C-BARQ in different studies (Henriksson 2016; Posluns et al. 2017). In a study conducted by Posluns et al. 2017 which included 60 dogs, the DPQ and the MCPQ-R scales (scored by dog owners(N = 60) and dog walkers(N = 10)) had considerable overlap. Inter-rater reliability studies were significant for three MCPQ-R scales; namely, Amicability, Neuroticism, and Extraversion; and three DPQ scales which were Aggression towards people, Aggression towards animals, and Fearfulness. The overlap between these two rating methods suggests a narrowing in the general trait taxonomy of canine personality. The results from this study suggested that the pairs Neuroticism (MCPQ-R) and Fearfulness (DPQ), Training focus (MCPQ-R) and Responsiveness to training (DPQ), and Extraversion (MCPQ-R) and Activity/Excitability (DPQ) are essentially identical. They also found Amicability (MCPQ-R) and Aggression towards people and animals (DPQ) measure the same construct. The MCPQ-R and the DPQ are sensitive to the different settings in which the assessor bases their evaluation of the dog. It is suggested that further research should be conducted to determine observer bias, especially as this behavioural assessment may have a direct effect on the animal's welfare such as working dogs (Posluns et al. 2017).

Rayment et al. 2016 created an online survey for pet dogs using a combination of the DIAS, C-BARQ and MCPQ-R. Completed surveys (N = 544) were used for analysis. Principal component analysis was performed on C-BARQ, DIAS and MCPQ-R. C-BARQ had a 13-factor solution, DIAS had a 3-factor solution and MCPQ-R had a 5-factor solution, reproducing the same results described by Ley et al. 2009a.

Statistically significant correlations were found between factor scores from different ratings. The MQPQ-R factor 'Neuroticism' had positive correlations with three C-BARQ fear-related factors, DIAS 'Response to Novelty' and a negative correlation with MCPQ-R 'Amicability'. The MCPQ-R 'Amicability' factor also showed negative correlations with many C-BARQ factors associated with fearful and aggressive behaviour, in addition to two measures of impulsivity. The MCPQ-R 'Training focus' factor had a moderate negative correlation with DIAS 'Behavioural regulation' factor and a moderate positive correlation with DIAS 'Responsiveness' factor. Another MCPQ-R factor, 'Extraversion' showed positive correlations with DIAS 'Behavioural regulation' factor and C-BARQ 'Excitability' factor. Many factors within the C-BARQ correlate with each other, the MCPQ-R had two inter-factor correlations and the DIAS showed no statistically significant correlations (Rayment et al. 2016).

Hunt et al. 2020 analysed the predictive validity of r-PWQ and C-BARQ in juvenile guide dogs (N = 359). The r-PWQ (completed by puppy raisers) had five predictive traits out of nine and the C-BARQ had two predictive traits out of eight (Hunt et al. 2020). This suggests that the r-PWQ is more suitable for working dog populations when compared to the C-BARQ. Cleghern et al. 2018 and Bray et al. 2019 have used C-BARQ rating scores and BCL/IFT test battery ethogram scores to predict training outcomes for prospective dogs in the guide dog programme. Cleghern et al. 2018 who conducted their study on guide dogs achieved an overall accuracy of 51% and 61%. Bray et al. 2019 who conducted their study on assistance dogs achieved 68% and 66%, respectively. When using a dataset of the same size, the C-BARQ performed better than the IFT/BCL. In both cases, the combination of behavioural scores from both sources into a dataset leads to a slightly poorer accuracy. However, it is important to highlight that the highest f1-score (measure of performance) was attained when taking only statistically significant variables, which included 5 C-BARQ scores and 22 IFT scores (Cleghern et al. 2018). Moreover, the difference was only marginal where the accuracy was 65% for the C-BARQ, 63% for the IFT and 64% for the combined dataset (Bray et al. 2019). Therefore, further research is needed to identify a more efficient way to combine scores from two assessment methods.

### 2.5.5 Canine Activity Monitors

Canine activity monitors are collar-based devices that use inertial sensors, such as accelerometers to measure physical activity in dogs. They are also referred to as smart collars or activity trackers. These devices are an inexpensive and reliable method for monitoring activity levels of canines in various different settings (Yashari et al. 2015).

Previous studies suggest a possible influence of physical activity on the behaviour of dogs, as it allows dogs to learn how to cope with frustration and maintain self-control (Zilocchi et al. 2016). A canine activity monitor could be beneficial in monitoring the time to complete the walking routes in assistance dog training programmes (Mizukoshi et al. 2008).

The possibility of estimating canine posture using IMUs was first investigated by Ribeiro et al. 2008 and Brugarolas et al. 2012 and subsequent attempts were introduced by Ribeiro et al. 2009; Brugarolas et al. 2013a; Brugarolas et al. 2013b. In terms of hardware design, posture recognition systems based on inertial sensing, as reported in the literature, either employed custom-designed devices (Bozkurt et al. 2014; Brugarolas et al. 2012; Brugarolas et al. 2013a) or commercially available devices (Ladha et al. 2013; Kiyohara et al. 2015; Kumpulainen et al. 2018; Ferdinandy et al. 2020; Kumpulainen et al. 2021). In terms of software development, these systems rely on classification algorithms which use statistical models (J. Majikes et al. 2017), supervised machine learning (Brugarolas et al. 2013a; Brugarolas et al. 2013b; Ladha et al. 2013) or knowledge engineering (Ribeiro et al. 2009) techniques to identify key canine postures and activities. Some canine activity monitoring devices are currently commercially available (Belda et al. 2018; Chambers et al. 2021; den Uijl et al. 2017; Marcato et al. 2022).

Table 2.3 summarises, in chronological order, nine published studies whose main goal was to develop a canine posture estimation system based on inertial sensors. It includes information on subjects, sensors and placement, data collection methods, data pre-processing techniques, classification algorithms, as well as train-test split, and results.

Table 2.3: Canine posture recognition algorithms using IMUs reported in literature.

| Publi-cation | Subjects<br>Sensors & Placement | Postures<br>Data Collection | Pre-processing<br>Classification Algorithm | Train-Test Split / Results |
|---|---|---|---|---|
| Ribeiro et al. 2009 | 5 Urban Search and Rescue dogs (breed not available)<br>Custom-designed system<br>2-axis ACC<br>Wither and rump | 3 static: standing, sitting, lying down 1 dynamic: walking<br>Postures repeated 5 times | Window size of 2 s and overlap is not reported<br>Calculation of angles from ACCs<br>Algorithm: custom, uses angles of ACCs | Training and test sets contain the same dogs from Ribeiro et al. 2008<br>Overall accuracy was 80% |
| Ladha et al. 2013 | 18 dogs (13 breeds)<br>AX3, Axivity<br>ACC @ 30Hz<br>Neck collar | 17 behaviours: bark, chew, dig, drink, eat, excrete, jump, lie down, sit, pawing, run, shake, shiver, sniff, urinate, walk, unspecified<br>Sessions lasted 20-40 minutes, freely moving dogs | Window size of 1 s and overlap of 0.5 s PCA-based feature extraction (50 components, 95% of variance) Normalisation using ECDF<br>Algorithm: K-nearest Neighbours (KNN) with k=1 | 10-fold stratified cross-validation<br>Overlap between training and test frames can be of 48%<br>Overall accuracy was 69% |

Continued on next page

Table 2.3 - continued from previous page

| Publi-cation | Subjects<br>Sensors &<br>Placement | Postures<br>Data Collection | Pre-processing<br>Classification Algorithm | Train-Test Split /<br>Results |
|---|---|---|---|---|
| Gerenc-sér<br>et al.<br>2013 | 24 dogs (12 BM, 12 LR)<br>Custom-designed system<br>ACC and GYR @ 100Hz<br>Back | 3 static: sit, lay, stand<br>4 dynamic: run, trot, walk, bark, search<br>Postures recorded on 2 days, sessions lasted 10min | Window size of 1 s and overlap is not reported<br>126 Features: 69 from GYR, 45 from ACC<br>Algorithm: Support Vector Machine (SVM)<br>Hyperparameter optimisation: soft margin (C) and gamma | 5-fold cross-validation, controlling for subjects<br>Single dog accuracy:<br>Intra-subject (BM 91%, LR 92%)<br>Intra-breed<br>Inter-subject (BM 70%, LR 73%)<br>Inter-breed (74%)<br>Multiple dog accuracy:<br>Inter-subject 83% |

Continued on next page

Table 2.3 - continued from previous page

| Publi-cation | Subjects<br>Sensors & Placement | Postures<br>Data Collection | Pre-processing<br>Classification Algorithm | Train-Test Split / Results |
|---|---|---|---|---|
| Bru-garolas et al. 2013a | 7 dogs (6 LR, 1 KK)<br>Custom-designed system<br>ACC and GYR @ 10Hz<br>Rump, Chest, Abdomen and Withers | 5 static: sitting, standing, lying down, eating off ground, standing on two legs<br>3 dynamic: walking, climb stairs, down ramp<br>Static postures repeated 5 times for 4s + climb stairs, down ramp | Window size of 1 s and overlap is not reported Moving average filter<br>Cascade machine learning algorithm<br>C1. separate static behaviour and classify dynamic behaviour<br>C2. posture classification for static behaviour | 5-fold cross-validation<br>Test set was 30% of data<br>Intra-subject accuracy:<br>C1. 92-100% C2. 78-99.5%<br>Inter-subject accuracy:<br>C1. 0-100% C2. NA |
| Bru-garolas et al. 2013b | 2 dogs (1 LR, 1 KK)<br>Custom-designed system<br>ACC and GYR @ 10 Hz<br>Neck and rump | 5 static: sit, stand, lie, eat off the ground, stand on two legs<br>Static postures repeated 5 times for 4s on 2 days | Window size and overlap are not reported Moving average filter<br>Algorithm: Random Forest (RF), KNN, and Logistic Model Tree (LMT)<br>Two-stage cascade learning technique used:<br>C1. separated postures and transitions (ACC, GYR) C2. classification of 5 postures (ACC) | 10-fold cross-validation<br>Inter-subject accuracy:<br>C1. 81 - 92% C2. 92 -100% |

Continued on next page

Table 2.3 - continued from previous page

| Publi-cation | Subjects<br>Sensors & Placement | Postures<br>Data Collection | Pre-processing<br>Classification Algorithm | Train-Test Split / Results |
|---|---|---|---|---|
| Kiy-ohara et al. 2015 | 2 dogs (Pembroke Welsh Corgi, Toy Poodle)<br>AX3, Axivity<br>ACC @ 25Hz<br>Neck Collar | 10 behaviours: walking, eating, sitting, laying, sniffing, running, jumping, drinking, shaking, scratching<br>No information available on data collection procedures | Window size of 1 s and overlap 0.96 s<br>Features from PCA and ECDF<br>Algorithm: K-nearest Neighbours (k=1)<br>Distance metric: Euclidean Distance (ED), Dynamic Time Warping (DTW), DTW-D proposed by Chen et al. 2013 | Training and test sets contain the same dogs<br>Distance metric DTW-D outperformed ED and DTW<br>Mean F1-measure 67% and accuracy is 78% |
| Kumpu-lainen et al. 2018 | 24 dogs (20 breeds)<br>GT9X Link, ActiGraph<br>ACC and GYR @ 100Hz<br>Neck collar | 7 behaviours: stand, sit, lie down, walk, trot, gallop, sniff<br>Each task lasted 3 minutes | Window size of 2 s and overlap is not reported<br>27 features created and Forward Feature Selection (FFS)<br>Algorithm: Linear and Quadratic Discriminant Analysis classifier (LDA and QDA) | 10-fold cross-validation<br>Train-test split criteria not available<br>Accuracies All: (27 features) LDA 74% and QDA 73% FFS: (8 feat) LDA 73% and (5 feat) QDA 76% |

Continued on next page

Table 2.3 - continued from previous page

| Publication | Subjects<br>Sensors & Placement | Postures<br>Data Collection | Pre-processing<br>Classification Algorithm | Train-Test Split / Results |
|---|---|---|---|---|
| Ferdinandy et al. 2020 | 21 dogs (19 breeds)<br>Apple Watch Series 1<br>ACC and GYR @ 50 Hz<br>Neck collar | 8 behaviours: sit, lay, stand, walk, trot, run, eat, drink<br>No information available on data collection procedures | Window size of 1.3 s and overlap of 0.8 s<br>252 features<br>Models created for small, medium and large sized dogs (N = 7 in each group)<br>Algorithm: SVM<br>Hyperparameter optimisation: soft margin (C) and gamma | Leave one subject out<br>Average accuracy 57% (results taken on comparable approach) |
| Kumpulainen et al. 2021 | 45 dogs (27 breeds)<br>GT9X Link, ActiGraph<br>ACC and GYR @ 100Hz<br>Neck and back | 3 static: sit, stand, lie down 4 dynamic: trotting, walk, play, treatsearch<br>Each task lasted 3 minutes | Window size of 2 s and overlap of 1s<br>54 features and FFS<br>Algorithm: LDA, QDA, SVM and Tree | Leave one subject out<br>Accuracy using the best classifier (SVM): Back 91.4% and Neck 75.6% |

Continued on next page

Breeds: LR = Labrador Retriever, BM = Belgian Malinois, KK = Kai Ken. Sensors: ACC = Accelerometer, GYR = Gyroscope. All sensors are 3-axis except stated otherwise. Features: PCA = Principal Component Analysis, ECDF = empirical cumulative density function.

Even though most of the studies reported overall accuracy as the main performance measure, comparisons across studies must be carefully made. This metric is highly influenced by class imbalance, in other words, accuracy becomes unrepresentative of the algorithm's performance as the number of observations in each category becomes dissimilar. Moreover, using data from the same subjects or same breeds in the training and test sets causes the performance metric to be higher compared to using different subjects (Brugarolas et al. 2013a; Ferdinandy et al. 2020) or breeds (Gerencsér et al. 2013). Controlling for these factors considering the final model's intended application is vital in estimating its real-world performance. Although the number of observations per class and subject as well as the criteria used for splitting the original dataset into training and test sets significantly affect the model's accuracy, several studies did not clearly report them.

Additionally, some studies lack information regarding important methodological details including the device sampling rate (Ribeiro et al. 2009), window size (Brugarolas et al. 2013b) and overlap (Ribeiro et al. 2009; Brugarolas et al. 2013a; Gerencsér et al. 2013; Kumpulainen et al. 2018), data collection methods (Kiyohara et al. 2015) and feature extraction methods (Ladha et al. 2013). Insufficient information and incompatible methodological designs pose great challenges when comparing the performance achieved in different studies and limit their contribution to state-of-the-art knowledge.

The number of subjects used for developing canine posture recognition algorithms was rather limited in early research; however, bigger sample sizes were reported in more recent studies (Gerencsér et al. 2013; Kumpulainen et al. 2018; Ferdinandy et al. 2020; Kumpulainen et al. 2021). Gerencsér et al. 2013 used two breeds of dogs and attained an inter-subject accuracy of 83% and Kumpulainen et al. 2018 used 20 breeds of dogs and achieved an accuracy of 76%. The lower figure in the latter study can be explained by the use of dogs from different breeds as it introduces more variability negatively impacting that metric. However, it is not clear whether the same subjects were used for training and testing, which would in turn positively affect that metric (Gerencsér et al. 2013; Ferdinandy et al. 2020).

In observational tests, L. M. Tomkins et al. 2011 found that apprentice guide dogs that spent a greater amount of time resting in the evening had a higher chance of being successful in their training programme than restless dogs. These results were obtained through a focal sample of 15-hour footage. They also used an Actical device to monitor the activity levels of the dogs in their kennels. Surprisingly, the activity levels measured using the Actical were not predictive of the dog's success. Modern canine activity monitors are capable of recording the time spent resting and making the data available on a smartphone application. It can be hypothesised that more meaningful results would

have been achieved if an appropriate canine activity monitor had been used or more data pre-processing techniques had been applied to the raw accelerometer data obtained from the Actical. This poses a great opportunity for researchers and organisations to collect a large amount of data with considerably less effort in order to analyse dog behaviour in a more integrative approach.

A study conducted on shelter dogs (N = 26) by Hiby et al. 2006 could have also benefited from the use of canine activity monitors if such technology were available then. Behaviour and urinary cortisol/creatine ratios (C/C) were monitored. Dogs were filmed for 30 minutes in their kennels where 21 behaviour patterns were observed. Four of these behaviours significantly changed over the ten days following their admission to a re-homing kennel. Drinking and grooming increased over time while panting and paw-lifting decreased. Dogs involved in the study who were observed drinking on their first day had significantly lower C/C ratios than those who were not observed drinking. Dogs with rising C/C ratios were on average more active than those with falling C/C ratios. This study suggests further investigation is required to examine the complex relationship between exercise and C/C ratios (Hiby et al. 2006). Canine activity monitors not only record activity in great detail, but some can also record drinking and grooming levels, such as Invoxia 2021 and Whistle 2021. This information may be beneficial to future research in this area.

### 2.5.6 Discussion on Behavioural Assessments

There is a range of behavioural assessment methods used to evaluate performance in trainee assistance dogs in the literature. They can be grouped into three categories; namely, test batteries, ratings, and observational tests. Additionally, canine activity monitors were identified as an emerging wearable technology that can potentially assist in the behavioural monitoring of working animals. Assistance dog organisations utilised a combination of these methods to assess prospective dogs at different stages of training.

Test batteries summarised in Table 2.1 were designed to assess the behaviour of prospective working dogs at different stages of development. A greater number of test batteries were applied to adult dogs than puppy or juvenile. Puppy testing is described as the 'holy grail' of behavioural assessment for the potential value in predicting future behaviour (Harvey et al. 2016b). It would be highly beneficial to the dogs, clients, and organisations if it were possible for host organisations to predict the future behaviour of their dogs as early as 12 weeks of age. Juvenile testing represents a great opportunity for future research as it is the least studied stage of development (Harvey et al. 2016b). However, evaluating dogs in their puppyhood or juvenile period

while with their puppy raisers presents additional challenges as their consent and continued participation are needed during the study (Alcaidinho et al. 2014).

It is proposed that 14 months of age is the optimal age for use of behavioural assessment methods (L. Batt et al. 2008b), which conveniently coincides with the age when dogs usually start their formal training. A common practice among some assistance dog organisations is to perform a behavioural assessment on prospective trainee dogs when they arrive at the training centre before starting formal training. This assessment can help determine whether the dog should proceed to a training programme, a breeding programme or be re-purposed as an ambassador, community, or pet dog. This is an important step in the training of assistance dogs in their respective organisations as training a dog that lacks the capabilities of becoming an assistance dog raises significant animal welfare concerns and causes a waste of resources. Hence, it is in the best interest of assistance dog associations to be able to identify unsuitable dogs to be withdrawn as early as possible (Mizukoshi et al. 2008).

As shown in Table 2.1, test batteries were developed to evaluate different aspects of behaviour; for instance, the IFT focuses on temperament traits, the DCTB focuses on cognitive abilities and the test battery created by L. Batt et al. 2008b is predictive of overall failure. Carrying out test batteries can be time-consuming and requires skilled assessors making it an expensive method of evaluating canine behaviour (Wiener et al. 2016). It is notable that as dog's age, the behaviour tests evolve. Puppy testing appears to mainly focus on measuring certain behaviours such as human interest, response to stimuli, and retrieval of objects. Juvenile testing also follows this format, focusing more on measuring behaviours such as obedience and distractions. Test batteries in adult testing focus more on the assessment of traits, along with specific behaviours. This may be due to the fact that performing a test battery on an adult dog is more successful in predicting the future behaviour of that dog (L. Batt et al. 2008b) as behavioural and social maturity is reached (Harvey 2021). While the lack of consistency in the stimuli used, age of testing, the form of evaluation and validation was previously criticised (Asher et al. 2013), it can be viewed as a necessary step that will allow for future successful standardisation of working dog assessment (Bray et al. 2021a).

Ratings are also a powerful tool in predicting the behaviour of assistance dogs (Bray et al. 2019; van Rooy et al. 2019). The use of ratings in dog organisations typically involves the completion of a questionnaire by puppy raisers, trainers or another domain expert who has spent a lot of time with the dog. Their experience and judgement are fundamental for the resulting scores to reflect a dog's general behaviour and typical response to a variety of situations (Bray et al. 2021a). The main

advantage associated with ratings is that they are easier, faster, cheaper (require less skilled assessors) and allow for a broader assessment of the dog's behaviour compared to the constrained sample evaluated in test batteries (van Rooy et al. 2019). Their main disadvantage is that they include a higher degree of subjectivity as a larger number of raters are involved, which yields communication challenges and low response rates (Bray et al. 2019). For this reason, some organisations prefer test batteries over ratings as they are free of owner bias and are evaluated by a trained observer who has a greater understanding of the normal and abnormal behaviours of dogs. However, as each method has advantages, many organisations use both (Wiener et al. 2016).

Some studies have used both a test battery and a rating when assessing the behaviour of assistance dogs (Cleghern et al. 2018; Dollion et al. 2019), personality (Barnard et al. 2016; Mirkó et al. 2013), and aggressiveness (Siracusa et al. 2010) in pet dogs, and shelter (Clay et al. 2020) and military dogs (Foyer et al. 2014; Lazarowski et al. 2021). Previous comparisons between the methods show that test batteries involving the rating of behaviours achieved higher reliability than the rating of individual dogs on broad traits to evaluate consistencies in animal behaviour (Harvey et al. 2016b; Harvey et al. 2017). Additionally, ratings of behaviours are less subjective and more practical (Robins et al. 2009). Barnard et al. 2016 reported results between a rating method and test battery to be rather consistent, as there were good correlations between them. However, Mirkó et al. 2013 suggests that more collaboration is needed between those developing test batteries and rating methods in order to increase correlations between them.

Observational tests are a more informal method of assessing dog behaviour allowing the interpreter to conclude with a broader understanding of the dog's behaviour than provided by a test battery (Amanda C. Jones et al. 2005). It involves analysing behaviour in a more naturalistic setting, for example during a walk (Murphy 1998; Mizukoshi et al. 2008) or when in the kennel (L. M. Tomkins et al. 2011). Observational tests can be completely informal or include some sub-tests that are similar to test batteries (Mizukoshi et al. 2008). Each of the observational tests mentioned originated significant variables suggesting that they are promising methods. Current literature focuses on test batteries and ratings; however, it may be beneficial for organisations to also employ observational tests as it is a more informal and simpler method which produces satisfactory results.

The use of canine activity monitors has become a recent trend and has been utilised in diverse areas of research such as sleep disturbances in brachycephalic dogs (Barker et al. 2021), monitoring activity to assess pruritus (Schwab-Richards et al. 2014), managing livestock guarding dogs (Zingaro et al. 2018), measuring physical activity (Colpoys et al. 2021; Yashari et al. 2015), and using inertial data to replace trained

scorers during behavioural assessments (Mealin et al. 2020). There is potential for their use in automating the behavioural assessment of dogs and predicting the training outcome of apprentice assistance dogs.

Dogs that were withdrawn due to behavioural reasons are one of the greatest challenges assistance dog organisations face. They are sometimes only deemed to be unsuitable for their role after undertaking a considerable amount of formal training and assessments. These personality traits are usually identified in an unsystematic manner based on behavioural cues shown in casual interactions or training sessions (J. A. Serpell et al. 2001). In an effort to increase efficiency and reduce cost, researchers and practitioners developed systems to capture behavioural information in a more objective manner through the assessment of canine temperament, personality, and cognitive abilities.

However, assessment systems reliant only on behavioural response are quite limited as they overlook important factors. Firstly, the scoring method depends on the raters' familiarity with the dog and experience in this area which introduces inter-rater variability. Secondly, the observable behavioural response depends on many other internal factors such as impulsivity, arousal, coping style (passive/reactive or active/proactive), affective state (positive or negative), and judgement bias (optimistic or pessimistic); and external factors like temporal and contextual stability. Moreover, each of the personality assessment models proposed has other limitations such as a lack of definition of the traits assessed, inconsistent terminology or incomplete personality structure. The employment of mixed behaviour assessment methods, such as ratings and test batteries, is recommended to mitigate some of these limitations (Rayment et al. 2015).

## 2.6 Physiological Assessment Methods

As discussed, assistance dog organisations utilised behavioural assessment methods for the early identification of personality traits associated with training outcomes. However, behavioural response (how one reacts) is affected by the affective state (how one feels) and cognition (how one interprets the stimuli) (Rayment et al. 2015). Affective states have two main dimensions, emotional valence (positive vs. negative) and arousal (intensity of activation). Researchers have stated those can be only meaningfully determined by analysing both behavioural and physiological responses (Beerda et al. 2000).

In the context of a trainee assistance dog, the dog's learning ability is crucial to the successful completion of the training programme. Learning ability is affected by factors such as stress, affective state, and levels of arousal which, in turn, have been demonstrated to correlate with physiological markers (Blackwell et al. 2010; Starling et al. 2013).

Published research revealed a significant association between physiological markers and training outcomes in military working dogs (Foyer et al. 2016). Furthermore, physiological signals are associated with undesirable behavioural traits like distractibility (Maros et al. 2008), impulsivity (Hannah F. Wright et al. 2012), fear (Hydbring-Sandberg 2004; Horváth et al. 2007), aggression (Wormald et al. 2016; Craig 2016; Craig et al. 2017; Horváth et al. 2007), and anxiety (R. C. M. M. Lensen et al. 2019; Wormald et al. 2017; Pirrone et al. 2019; Bray et al. 2017b), as well as desirable traits like trainability (R. C. M. M. Lensen et al. 2019), problem-solving (Blackwell et al. 2010), and adaptive ability (Kikkawa et al. 2005; Bray et al. 2017b) which affect the probability of successful completion in assistance dog programmes.

Physiological and behavioural responses do not always show correlation as they are differently influenced at an individual level by their coping style (J. Koolhaas et al. 1999) and current affective state (Bonne Beerda et al. 1998). For instance, activity influences HR only in the absence of emotional stimuli, suggesting that physiological and behavioural responses represent different dimensions (Palestrini et al. 2005). Therefore, if only behaviour is analysed the other dimension is lost. The link between physiology and observable behaviour is of great importance in expanding the current understanding of the dog's emotional reactions to environmental stimuli.

Accordingly, there is a growing body of research exploring physiological signals in the field of canine performance assessment. This section explores the opportunity presented by the inclusion of physiological monitoring to complement behavioural analysis in quantifying canine emotional response, cognitive abilities, and personality traits. This multi-parameter evaluation can be the key to advancing the current practices in estimating trainee dogs' suitability for assistance work.

The following subsections will cover (1) a brief introduction of relevant physiological systems involved in the emotional response; (2) a review of established physiological parameters used to quantify emotional or affective state (emotional valence and arousal); and (3) an overview of commercially available physiological devices as well as important considerations and practical advice on the criteria for selection of such devices for canine performance assessment.

### 2.6.1 Systems

The stress response system comprises the autonomic nervous system (ANS) and the hypothalamic pituitary adrenal (HPA) axis (Turner et al. 2020). The study of the correlations between the ANS activity and affective state (emotions) was first introduced almost a century ago (Cannon 1916). Since then, this topic has been vastly explored in the

psycho-physiology research area.

The autonomic nervous system (ANS) is systematically divided into two branches; namely, the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). The SNS is often referred to as the fight-or-flight system while the PNS is known as rest-and-digest. PNS activity is also called vagal tone as the vagus nerve is a fundamental component of the PNS. The relationship between SNS and PNS activity is referred to as sympatho-vagal balance.

Autonomic or sympatho-vagal balance reflects the state of the ANS considering the opposite effect of its components - SNS and PNS activity. Sympatho-vagal balance reflects the individual's ability to alter the relative activity of the SNS and PNS. While both PNS and SNS are active at any given time, some individuals are prone to either SNS or PNS dominance (Berntson et al. 2008). An efficient autonomic activity is linked with PNS dominance also known as high vagal tone (Brodal 2003). The increased vagal tone reflects a higher regulatory sensitivity and flexibility translating into an ability to respond appropriately to physiological or environmental stressors (Borell et al. 2007). Conversely, a decreased vagal tone indicates an inability of the autonomic system to cope with external stimuli and is associated with numerous clinical conditions and behavioural issues (Berntson et al. 2008).

The ANS controls the functioning of the cardiovascular system, therefore its activity can be measured through HR, HRV, and Respiratory Sinus Arrhythmia (RSA) (Grossman et al. 2007). Two main neurohormonal pathways are activated in response to stressors. The endocrine system activates the HPA axis and the nervous system activates the sympatho-adrenal medullary (SAM) axis both in humans (Godoy et al. 2018) and dogs (Kooriyama et al. 2021). The HPA axis produces long-term responses to stress stimuli and its activation results in the secretion of glucocorticoid hormones, in dogs the most predominant one is cortisol. Cortisol has an immunosuppressive effect, therefore the immune biomarker called salivary secretory immunoglobulin A (sIgA) was investigated as an indirect indicator of stress response in dogs (Skandakumar et al. 1995; Kikkawa et al. 2003). The SAM axis provides a short-term response to stress stimuli and its activation leads to a release of catecholamines (adrenaline and noradrenaline) which stimulate the circulatory, cardiovascular, and respiratory systems increasing blood pressure (BP), Heart Rate (HR), and Breathing Rate (BR), respectively (Godoy et al. 2018).

### 2.6.2 Parameters

Kreibig 2010 conducted a search in 134 published research papers to investigate the most popular physiological measures used to investigate

ANS activity in response to emotions. The following measures of cardiovascular, respiratory, and electrodermal activity were traditionally examined in the literature, in order of frequency: Heart Rate (HR), Skin Conductance Level (SCL) / Electrodermal Activity (EDA), Blood Pressure (BP) - Diastolic BP and Systolic BP, Heart Rate Variability (HRV), Breathing Rate (BR), Finger Temperature (FT), Respiratory Sinus Arrhythmia (RSA).

In dogs, cortisol (Bonne Beerda et al. 1998; Beerda et al. 2000), tachycardia (rapid heartbeat) and tachypnoea (rapid respiration) (Bodnariu 2008) are physiological signs of stress. Although there is no established gold standard for measuring emotional state, cortisol, catecholamines, and behaviour, they are the most common parameters observed in dogs (Bodnariu 2008).

In recent studies, the correlation between changes in HR, HRV, BR and skin temperature with the emotional state were studied. Hence, the following physiological measures of emotion and temperament in dogs are reviewed: cardiac (Heart Rate, Heart Rate Variability) and respiratory (Breathing Rate). Moreover, Appendix A discusses the use of the following hormonal indicators: cortisol, catecholamines or chromogranin A, and salivary Immunoglobulin A.

#### 2.6.2.1 Cardiac

The cardiac system is controlled by the SNS and PNS in an independent inverse fashion: the sympathetic system increases cardiac activity and the parasympathetic decreases cardiac activity (Bernston et al. 1997). Heart Rate Variability (HRV) is the number of heartbeats in one minute. It is a measure of cardiac activity modulated by SNS and PNS. HRV has been included in previous research in canine behaviour assessment to supplement behavioural analysis yielding a more complete understanding of the cognitive and affective reaction to different stimuli (Palestrini et al. 2005).

For instance, guide dogs showed a deceleration in HRV during an attentional state when gathering information on the environment (Fallani et al. 2007). During a period of distress, dogs had a significantly high HRV especially if they were classified as fearful of that particular situation (Hydbring-Sandberg 2004) even if their activity levels were low (Palestrini et al. 2005; Fallani et al. 2007). These findings indicate that HRV only represents the level of arousal and does not allow for distinction between positive and negative emotional valences (R. C. M. M. Lensen et al. 2017; Gácsi et al. 2013).

Heart Rate Variability (HRV) parameters are indicators of cardiac activity and are an important measure of ANS activity. They are influenced by the activity and balance of the SNS and PNS, therefore

representing the regulatory capacity of the ANS (Shaffer et al. 2017). HRV parameters are metrics derived from the time interval between heartbeats called Interbeat Interval (IBI), also known as R-R (R-to-R or RR) intervals, as the heartbeats essentially are the R peak in an electrocardiogram (ECG). IBI or R-R intervals are filtered to remove artefacts producing the so-called NN intervals which are defined as the time between adjacent normal heartbeats. HRV metrics are computed by applying time and frequency-domain statistical measures on the NN measurements over a period of time.

Commonly used time-domain metrics include the following: standard deviation of IBI (SDRR), average of NN intervals (AVNN), standard deviation of NN intervals (SDNN), standard deviation of AVNN (SDANN), average of SDNN (SDNNI), number of adjacent NN intervals that differ more than 50 ms (NN50), percent of adjacent NN intervals that differ more than 50 ms (pNN50), root mean square of successive differences of NN intervals (rMSSD), and average difference between the highest and lowest HRs during a respiratory cycle (HR Max - HR Min), integral of the density of the RR interval histogram in 24h divided by its height is called HRV Triangular index (HTI), triangular interpolation of the NN interval Histogram (TINN). Popular frequency-domain parameters include total power (TP), ultra-low frequency power (ULF), very low-frequency power (VLF), low-frequency power (LF), high-frequency power (HF), LF/HF ratio, normalised LF (LFnu), and normalised HF (HFnu). Time-domain and frequency-domain parameters of HRV represent the activity of sympathetic and parasympathetic nervous, accordingly. Conventional non-linear metrics include S, SD1, SD2, and SD1/SD2 which are derived from Poincaré plots.

Various demographic, physiological, methodological, and contextual factors influence HRV metrics in complex ways. They should be taken into account in the interpretation of HRV, important considerations include the calculation period, IBI detection method, sampling frequency, removal of artefacts, respiration and activity levels have on heart rate measures when using it to assess emotional response. Some metrics are more sensitive than others to artefacts in the NN intervals (Shaffer et al. 2017).

HRV metrics are extensively discussed in the literature in relation to their psycho-physiological significance and pairwise relationships between different HRV metrics. While consensus is evident in a number of aspects, controversy prevails (Bernston et al. 1997). For instance, PNS activity or vagal tone can be measured through RMSSD and HF (Kleiger et al. 2005). On the other hand, researchers have not reached a consensus in relation to the validity of LF/HF ratio to measure sympatho-vagal balance (Shaffer et al. 2017). There are five different theories implying the use of HRV in psycho-physiological research, the common ground between them regards the focus on vagal tone. Vagal

tone influences the sympatho-vagal balance and their interaction can be measured through HRV parameters. Higher vagal tone is associated with better cognitive performance, physical (physiological) and mental (psychological) health (Thayer et al. 2009), social functioning (Porges 2007), and adaptive capability (Grossman et al. 2007). HRV metrics are linked with self-regulation capacity, which represents an individual's ability to cope with environmental stressors, indicating emotional stability and potential behavioural problems (Reynard et al. 2011; McCraty et al. 2015; Laborde et al. 2017).

HRV metrics are validated in research as a measure of stress in humans (McCraty et al. 2015) and welfare in farm animals (Borell et al. 2007), as well as behavioural issues (Wormald et al. 2016; Wormald et al. 2017; Craig et al. 2017) and emotions in pet dogs (Katayama et al. 2016; Zupan et al. 2016). They are used to estimate emotions and autonomic regulation in humans (Kreibig 2010; McCraty et al. 2009) and animals (Pagani et al. 1986). In fact, estimating canine emotional state based on HRV parameters was proved possible in past studies (Nakahara et al. 2016; Katayama et al. 2016; Zupan et al. 2016).

These preliminary results indicate a promising opportunity to apply cardiac monitoring in the particular context of trainee assistance dog performance and suitability assessment. HRV metrics present significant differences in groups of dogs affected by behavioural problems which influence their suitability as assistance dogs such as aggression (Wormald et al. 2016; Craig et al. 2017) and general anxiety-related issues (Wormald et al. 2017). Moreover, ECG data were successfully used to predict behavioural scores resulting from analysis of a guide dog assessment test battery (Mealin et al. 2020).

Respiratory measures are important in the interpretation of cardiovascular activity as both are modulated by the ANS through complex respiratory-circulatory interactions. The respiratory rate affects ECG and therefore HRV metrics; in particular, it influences frequency-domain measures such as LF, HF, total spectral power and any ratio between spectral bands (LF/HF) (Borell et al. 2007). Researchers highlighted the importance of establishing an appropriate baseline and monitoring respiration in psycho-physiological research involving HRV parameters (Quintana et al. 2014; Laborde et al. 2017). In dogs, respiratory rate has positive correlation with HF-HRV (Zupan et al. 2016). Therefore, it is recommended that researchers include respiratory monitoring in their data collection protocol when evaluating HRV metrics.

#### 2.6.2.2 Respiratory

Breathing requires complex interactions of the central and peripheral nervous systems, it is particularly influenced by the functioning of the

ANS and SAM axis. The most common measure of respiratory system activity is breathing or Breathing Rate (BR). In humans, the effect of emotional response on respiratory patterns is associated with task involvement and could be used to distinguish between passive/reactive versus active/proactive coping styles (Boiten 1998).

Increased BR (rapid respiration, panting) is a physiological indicator of acute stress (Bonne Beerda et al. 1997) and low BR (deep respiration, sighing/yawning) has been associated with a relaxed, positive emotional state (Katayama et al. 2016). Higher respiratory rate was reported as a sign of stress after long transport in animal-assisted intervention dogs (Carvalho et al. 2020), pain after surgery (Aengwanich et al. 2019; Srithunyarat et al. 2016; Sousa et al. 2017), and distress caused by noise in pet dogs (Jeong et al. 2020; Stellato et al. 2019).

This research indicates that respiratory rate is significantly affected by emotional state in dogs, therefore it can be used to improve the quantification of the response to stimuli presented in test batteries and observational tests. In fact, the occurrence of panting significantly reduced the chances of success in a guide dog training programme (L. M. Tomkins et al. 2011).

There are important interactions of the respiratory system with the cardiovascular system, as attested by the phenomenon of respiratory sinus arrhythmia (RSA) (Grossman et al. 2007). RSA was used as an indicator of emotional regulation capacity in humans (Beauchaine 2015).

### 2.6.3 Devices

There are efforts to design canine physiological monitoring devices for measuring heart rate and respiratory rate; however, some limitations must be overcome before the devices can be widely deployed and become commercially available (Foster et al. 2020; Antink et al. 2019). In this section, a selection of commercially available physiological measurement devices is discussed considering their suitability for psycho-physiological research on dogs. These devices were found through Google searches (using keywords such as heart rate monitor, cardiac monitor, ECG monitor, respiratory rate monitor, and physiological devices) and mentioned in research papers.

The physical size and shape of the device - also referred to as form factor - is a fundamental aspect in this particular application, as the main goal is to further understand the dog's affective state or emotional state in response to stimuli. Therefore, an essential requirement is that the device does not influence the subject by constraining their movements or eliciting emotions such as discomfort or stress.

Accordingly, collars are the ideal form factor for a canine physiological monitoring system due to their familiarity with such accessories. They also present significant advantages over implantable devices as they are minimally invasive and can be used on multiple subjects. The only smart collar currently available that is capable of measuring heart rate and respiratory rate is PetPace. Voyce used to offer a collar featuring similar capabilities and is working on a newer version. However, both devices have critical limitations in terms of reliability, accuracy, sampling rate, and data export options.

The gold standard for cardiac activity monitoring is traditional electrocardiogram (ECG) holter monitors which are coupled with disposable wet electrodes. They rely on good skin contact which is promoted by skin preparation including shaving, cleaning, and application of electrode conductive gel. These requirements pose several challenges, therefore there are recent effortsto design canine cardiac monitoring systems focusing on improved usability through the employment of dry electrodes (Brugarolas et al. 2015; Virtanen et al. 2018) and the use of methods other than ECG, such as optical (Brugarolas et al. 2014), radio (Landis-Hanna et al. 2016), and acoustic technology (Cotur et al. 2020; Belda et al. 2018; Ortmeyer et al. 2018).

Electrode belts used as chest straps offer an excellent compromise between the comfort of collars and the accuracy of ECG monitors. Essner et al. 2013 reported that dogs were very accepting of the Polar electrode belt, therefore it can be inferred that other chest strap devices would not cause any distress that could interfere with the physiological measurements. Therefore, research was carried out to identify devices suitable for monitoring cardiac and respiratory parameters in freely moving dogs.

Table 2.4 presents commercially available chest straps capable of monitoring a range of physiological signals that can indicate emotional state. Physiological devices featured meet the following inclusion criteria: wearable wireless devices featuring ECG-based heart rate monitoring, with the physical form of a chest strap, ability to measure IBI and raw data export options. Multi-parameter systems such as Bioharness and eq02+ Lifemonitor are a lot more expensive, have a significantly shorter battery life and use rechargeable batteries rather than disposable ones used in systems that only monitor cardiac parameters like Polar H10, Wahoo TICKR X, HRM-Pro Sensor, and Suunto Smart Sensor. Another important factor for research purposes is the method of accessibility to raw data since some of the devices require the use of specific paid software as in the case of Bioharness 3, eq02+ LifeMonitor, and Polar H10.

Table 2.4: Multi-parameter physiological devices and cardiac monitors physically composed of a sensing module attached to a chest strap.

| Device<br>Manufacturer<br>Publication | Parameters | Cost<br>Battery<br>Connectivity |
|---|---|---|
| Bioharness 3<br>Zephyr Technology, Annapolis, Maryland, USA<br>(Zephyr 2021)<br>(Hailstone et al. 2011; Johnstone et al. 2012a; Johnstone et al. 2012c; Johnstone et al. 2012b; Kim et al. 2013; Nepi et al. 2016) | ECG<br>IBI/Heart Rate Variability<br>Heart Rate<br>Breathing Rate<br>3-axis Accelerometer | € 1,865.99<br>12-35h, NA<br>Bluetooth, Zephy<br>ECHO Gateway |
| eq02+ LifeMonitor<br>Equivital Inc., Cambridge, UK<br>(Equivital 2021)<br>(Akintola et al. 2016; Liu et al. 2013) | ECG<br>Hear Rate<br>IBI/Heart Rate Variability<br>Respiratory Rate<br>Skin Temperature<br>Accelerometer | € 1,413.65<br>48h, Lithium Polymer<br>Bluetooth |
| HRM-Pro Sensor<br>Garmin, Olathe, Kansas, United States<br>(Garmin et al. 2021) | Heart Rate<br>IBI/Heart Rate Variability<br>Activity Monitoring | € 129.99<br>365h, CR2032<br>Bluetooth, ANT+ |

Continued on next page.

Table 2.4 - continued from previous page.

| Device<br>Manufacturer<br>Publication | Parameters | Cost<br>Battery<br>Connectivity |
|---|---|---|
| Polar H10<br>Polar Electro Oy, Kempele, Finland<br>(Polar 2021)<br>(Gilgen-Ammann et al. 2019; Carballo et al. 2020; Orozco et al. 2021; H. S. Lee et al. 2020; Hae Sung Lee et al. 2021; Harvie et al. 2021) | Heart Rate<br>IBI/Heart Rate Variability<br>Activity Monitoring | € 89.90<br>400h, CR2025<br>Bluetooth, ANT+ |
| Wahoo TICKR X<br>Wahoo Fitness, Atlanta, Georgia, USA<br>(Wahoo 2021)<br>(Polar 2019) | Heart Rate<br>IBI/Heart Rate Variability<br>Activity Monitoring | € 79.99<br>500h, CR2032<br>Bluetooth |
| Suunto Smart Sensor<br>Suunto Inc., Vantaa, Finland<br>(Suunto 2021)<br>(Weippert et al. 2010; Polar 2019) | Heart Rate<br>IBI/Heart Rate Variability<br>Activity Monitoring | € 79.00<br>500h, CR2025<br>Bluetooth |
| Cardiosport TP5 / Ithlete<br>Healthcare Technology, Hampshire, United Kingdom<br>(Cardiosport 2021; Ithlete 2021)<br>(Flatt et al. 2013) | Heart Rate<br>IBI/Heart Rate Variability<br>3-axis Accelerometer | € 60.00<br>900h, CR2032<br>Bluetooth, ANT+ |

eq02+ LifeMonitor was evaluated for monitoring multiple physiological parameters (Liu et al. 2013; Akintola et al. 2016) in humans and has been widely deployed in many research areas requiring ambulatory monitoring such as psycho-physiology as an indicator of emotions (Khan 2011; Agrafioti et al. 2012) and well-being (Vollmer-Conna et al. 2020).

The next subsections review three devices that could be considered for monitoring physiological signals in dogs; namely, Polar H10 and Bioharness 3. Their advantages and limitations will be critically reviewed considering their application in psycho-physiological research on dogs.

#### 2.6.3.1 Polar H10

Perhaps the most common Heart Rate Monitors (HRMs) in the market are produced by Polar®(Polar Electro Öy, Kempele, Finland), their popularity is justified by their low cost and high accuracy (Polar 2019). Polar devices provide a compact solution for sensing cardiac activity, given that they could substitute ECG monitors when the application only requires IBI measurements. Researchers have validated previous devices for measuring IBI in humans and HR in dogs under different situations.

Figure 2.1: Polar H10 heart rate sensor.

Polar RS800CX with W.I.N.D sensor was validated for HR measurements in dogs under sub-maximal cardiovascular conditions - standing position and at trotting on treadmill (Essner et al. 2013). The effect of the IBI measurement error, given by the difference between ECG and Polar measurement, on HRV parameters was then studied by Jonckheer-Sheehy et al. 2012; Essner et al. 2015a. While in the first study, no significant differences in HRV were found, the second study revealed that even though the IBI series contained less than 5% of erroneous data, it negatively affected the time-based HRV parameters. It is important to consider that an old model (RS800CX + W.I.N.D) was used by Essner et al. 2013; Essner et al. 2015a; Essner et al. 2015b; however, a newer model Polar H7 was proven to be accurate enough to measure HRV (Giles et al. 2015) in humans during supine and standing positions.

Polar released the H10 model in 2017 claiming it to be the most accurate

sensor thus far, featuring extra interference-preventing electrodes and an entirely new measurement algorithm (Polar 2017). Polar H10 provides more accurate IBI data than ECG holters in humans while performing intense physical activity (Polar 2019), being considered the gold standard for IBI assessment even in conditions featuring intense activity (Gilgen-Ammann et al. 2019). Moreover, it was used to validate other heart rate sensors in humans (Müller et al. 2019; Olstad et al. 2020; Muggeridge et al. 2021).

In summary, the first generations of Polar HRM devices were not accurate enough to provide satisfactory IBI measurements for HRV calculation. The more recent Polar H10 device is considered a gold standard for IBI measurements, providing HRV comparable to those calculated based on ECG measurements. Considering the satisfactory use of older Polar devices on dogs and the outstanding performance of Polar H10 on humans for IBI measurements, it is possible to conclude that Polar H10 is a promising alternative to ECG holters for accurate HR and HRV calculation on dogs. Nonetheless, further research is needed to assess their accuracy on freely moving dogs as artefacts are usually attributed to poor skin contact resulting from movement.

#### 2.6.3.2 Bioharness

Zephyr Bioharness 3 (Zephyr Technology Ltd., Auckland, New Zealand) is another physiological monitoring device composed of chest strap electrodes to measure ECG signals, HR, HRV, BR and Activity. Although Bioharness has only been used on goats (Briefer et al. 2015b) and horses (Briefer et al. 2015a), because its form factor is very similar to that of Polar, it could be safely presumed that it will also be well accepted by dogs.

Figure 2.2: Bioharness, Zephyr

Bioharness™ BR measurements were compared to Metamax 3b (manufactured by Cortex (Cortex, Leipzig, Germany) (Hailstone et al. 2011). It was concluded that Bioharness™ is a valid and reliable device for determining Respiratory Rates during exercises of varying intensity in humans.

A second study compared HR and BR in humans obtained from the

Bioharness™ with those from Model K4 b2 (COSMED, Rome, Italy) during a graded exercise test and Vmax Spectra System (VIASYS, Yorba Linda, CA) during a heat exposure test. The measurements taken with both devices were highly correlated, therefore Bioharness™ was found to be reasonably accurate for HR and BR measurements. (Kim et al. 2013)

### 2.6.4 Discussion on Physiological Assessments

Physiological assessment of the ANS is a promising method for the evaluation of mental and physical health. The monitoring of multiple physiological parameters is recommended for a more accurate estimation of affective state (Hekman et al. 2014; R. C. M. M. Lensen et al. 2019). The main physiological parameters investigated in the literature on canine welfare and personality are hormonal indicators such as cortisol, secretory immunoglobulin A (sIgA), catecholamines, and chromogranin A (CgA), as well as cardio-respiratory signals such as Heart Rate (HR), Heart Rate Variability (HRV), and Breathing Rate (BR).

Diverse physiological biomarkers including cardiac, respiratory, and hormonal measures were investigated in research relating to stress emotional state (Katayama et al. 2016; Zupan et al. 2016; Bonne Beerda et al. 1998; Gácsi et al. 2013; Fallani et al. 2007; Palestrini et al. 2005), personality traits (Wormald et al. 2016; Wormald et al. 2017; Craig 2016; Craig et al. 2017; Hannah F. Wright et al. 2012; Ryan et al. 2019), coping styles (J. Koolhaas et al. 1999; Horváth et al. 2007; Blackwell et al. 2010), temperament and problem-solving test evaluation (Mealin et al. 2019; Bray et al. 2017a) and training outcome assessment (L. M. Tomkins et al. 2011; Kikkawa et al. 2005; L. Batt et al. 2008a). These preliminary results suggest a promising opportunity for the employment of physiological biomarkers to evaluate prospective working dogs.

Test battery protocols aiming to identify personality or temperament traits should record both the behavioural response expressed by the dog (through the deployment of behavioural scoring systems) and the affective state, also known as an emotional state (through the monitoring of physiological biomarkers), experienced by the dog. Similarly, we recommended that researchers include physiological assessment when performing test batteries and observational tests with the purpose of evaluating assistance dogs' suitability for work. The physiological measurements can be used along with the existing behavioural evaluation when analysing differences between successful and unsuccessful dogs. This approach would render a more complete estimation of the trainee dog's behavioural and emotional reaction to different stimuli presented on such occasions and how that relates to their training outcome.

There is no commercially available device with the capability of recording physiological health parameters accurately on dogs. Collars are the ideal form factor for long-term use on dogs; however, chest straps have been validated for short-term use. The current options present a significant compromise between suitable form factors and accuracy/reliability. For instance, the collar PetPace does not provide accurate and reliable measures of HR, HRV, and BR (PetPace 2021). At the time of this review, the chest strap Polar H10 seems to be the most suitable device for monitoring cardiac activity as it is the most accurate device for measuring IBI data for the calculation HR and HRV; nonetheless, it does not monitor BR.

Finally, more research is needed to validate commercially available devices and further explore physiological parameters' potential in predicting training outcomes in assistance dogs. We strongly recommend that manufacturers of physiological measurement devices or researchers validate measurements against a gold standard device and develop software to allow users to access raw data. We suggest that researchers deploy physiological devices when performing test batteries and observational tests to evaluate prospective assistance dogs' suitability for work.

## 2.7 Discussion

Conclusively, this chapter provides an extensive review of the literature on behavioural and physiological tools and devices for the selection and assessment of assistance dogs. It was partially published by Elsevier in the Applied Animal Behaviour Science journal (Marcato et al. 2022). The techniques reviewed ranged in a scale of novelty from the well-established behavioural assessment tools - namely, ratings, test batteries, and observational tests - to emerging techniques deploying novel technology, such as canine activity monitors and wearable devices, which measure physiological parameters such as cardiac and respiratory activity. This section introduces the research conducted in this thesis to address the gaps in the state-of-the-art behavioural and physiological methods for assessing a trainee dog's suitability for work.

The employment of multiple behavioural assessment methods such as test batteries, ratings, and observational tests is recommended to minimise temporal or contextual effects on the assessment results. The most popular standardised dog behaviour rating reported in the literature for predicting training outcomes in assistance dogs is the Canine Behavioral Assessment and Research Questionnaire (C-BARQ). Its main downside is that it contains 100 context-based items, which makes it unsuitable for dog trainers to answer due to the number and specific nature of the questions. Chapter 3 investigates for the first time

the use of the MCPQ-R to predict training outcomes in trainee dogs and the results were compared with the widely used C-BARQ. No study was found to utilise test batteries to assess the behaviour of trainee assistance dogs at different stages of training with the goal of predicting working outcomes. In Chapter 4, a novel test battery was developed utilising best practices to estimate working outcomes in trainee assistance dogs at different points of the training programme.

Commercially available canine activity monitors represent a great opportunity for evaluating the relationship between long-term behavioural patterns and training outcomes. These devices contain posture recognition algorithms capable of automatically identifying a set of canine behaviours. However, accurate predictions are critical for the application in this industry. In Chapter 5, a new posture recognition system was developed for assistance dogs with the goal of outperforming the state-of-the-art by combining improvements in hardware composition and software development. In particular, it incorporated multiple inertial sensors on a harness and applied advanced machine learning methods for the first time in this field. In order to facilitate and accelerate the process of benchmarking, validating and developing posture estimation algorithms, the annotated posture datasets will be made publicly available.

Interdisciplinary collaboration is strongly advised as a means to improve the evaluation of prospective assistance dogs' suitability for work. Acquiring both behavioural and physiological data allows for a finer understanding of the influence of coping strategies, emotional regulation, adaptive capacity, and personality traits on training performance (Mia Cobb et al. 2015). Limited research is available in the emerging field of physiological assessment methods for assistance dogs. This thesis contributes to this area by sharing physiological data collected during the test battery. In particular, two devices identified in this literature review were utilised for recording cardio-respiratory parameters. This approach reduces subjectivity, increases the validity, and expands current state-of-the-art knowledge about the relationship between behaviour and physiological response to stressful stimuli (Rayment et al. 2015).

# Chapter 3

# Ratings

## 3.1 Introduction

The work presented in this chapter aims to achieve the primary objective of this thesis which is to evaluate and compare the validity of traditional and novel behaviour assessment methods to predict training outcomes for trainee assistance dogs. As discussed in Chapter 2, a variety of studies have been carried out with the goal of assessing the suitability of dogs for assistance work. Behavioural assessment methods can be classified into ratings, test batteries, observational tests, and canine activity monitoring. This chapter discusses the use of different ratings for evaluating the suitability of trainee dogs for assistance work. In particular, the ratings of interest were canine behaviour and personality assessment questionnaires. These are convenient methods for dog training organisations to gather information about prospective and trainee dogs in an objective and systemic way as discussed in Chapter 2.

Two criteria were used in this research to identify possible ratings: (1) designed to assess all aspects of dog personality and behaviour, and (2) standardised and tested for reliability and validity purposes in previous studies. The following aspects were considered to select the most appropriate rating for the two groups of participants (puppy raisers and dog trainers): length, vocabulary, and familiarity. The length of the questionnaire affects the time it would take for participants to complete it. The vocabulary influences their understanding of the concept that the item aims to evaluate. Finally, familiarity considers their ability to answer the items based on previous experiences with the dog.

The Canine Behavioral Assessment and Research Questionnaire (C-BARQ) (J. A. Serpell et al. 2001; Hsu et al. 2003; Duffy et al. 2012) was deemed to be the most appropriate for the puppy raisers as they may not have sufficient understanding of the technical vocabulary used in

adjective-based ratings. Puppy raisers are still willing to complete this long questionnaire as they only answer it for one dog. The context-based items are not an issue in this instance, as puppy raisers are very familiar with the dog's responses in a range of situations.

The Monash Canine Personality Questionnaire - Revised (MCPQ-R) (Ley et al. 2009a; Ley et al. 2009b) was chosen for the dog trainers because they are not familiar with the dog's response to some particular situations referenced in context-based ratings. They usually complete the questionnaire for multiple dogs at a time; therefore, shorter rating questionnaires are preferred to promote participation. Although the vocabulary used in this adjective-based questionnaire can be considered technical, dog trainers usually have a great understanding of dog behaviour and personality traits.

The validity and reliability of both the C-BARQ and MCPQ-R were comprehensively evaluated in the literature. The validity and internal consistency of the C-BARQ and MCPQ-R were demonstrated in Hsu et al. 2003; Ley et al. 2008 and Ley et al. 2009a, respectively. The inter-rater and test-retest reliability of the C-BARQ and MCPQ-R were reported in Duffy et al. 2008 and Ley et al. 2009a, respectively.

The main goal of this study was to apply various machine learning methods to improve the prediction of training outcomes in trainee assistance dogs using ratings. In order to achieve that goal, two machine learning models were built separately using data from the C-BARQ as rated by puppy raisers, and MCPQ-R as rated by dog trainers. This is the first study to explore the employment of the MCPQ-R rating for this purpose; hence, it was also of interest to compare model performance with the widely used C-BARQ. The focus of this chapter is to investigate the association between behaviour and personality (as assessed using standardised ratings) and training outcomes. More specifically, it was hypothesised that these ratings can be used to differentiate between dogs that successfully completed the assistance dog training programme and those that were withdrawn for behavioural reasons. Hence, the following research questions were investigated:

**Q.1** Is there a significant association between C-BARQ and MCPQ-R ratings and training outcomes?

**Q.2** Can statistical models estimate training outcomes based on C-BARQ and MCPQ-R ratings?

**Q.3** Which is more predictive of training outcomes: the C-BARQ or the MCPQ-R?

## 3.2 Methodology

In order to achieve the aims proposed and answer the research questions presented, the IGDB invited puppy raisers to complete the rating C-BARQ when the dogs were due to be returned to their training facility for formal training at the age of 13 to 18 months. Subsequently, dog trainers were invited to complete the rating MCPQ-R when the dogs reached 10 weeks of formal training around the age of 15 to 21 months. The C-BARQ and MCPQ-R are introduced in Sections 3.2.3 and 3.2.4, respectively. Section 3.2.5 introduces the methodology employed to analyse the data collected.

### 3.2.1 Ethical Approval

The Social Research Ethics Committee (SREC) at University College Cork (UCC) reviewed and approved the questionnaire data collection procedures as described below under request number 2019-016. Puppy raisers, dog trainers, and dog instructors related to the dogs participating in the research reviewed the information sheet and signed the consent form before completing the questionnaires.

### 3.2.2 Subjects

All dogs starting formal training at the IGDB from April/2019 to April/2021 were recruited to participate in this study. Data collection was affected by the global COVID-19 pandemic resulting in the number of dogs starting training being much reduced. Nevertheless, invitations to complete the questionnaires continued to be sent by email to puppy raisers and dog trainers as usual because the whole process was carried out remotely.

Dogs who completed the training were classified as “success” after qualifying as service or guide dogs and being matched with a blind or disabled person, respectively. Dogs withdrawn from the training programme for behavioural reasons were classified as “fail”. Dogs who were withdrawn from the training programme due to health issues or selection as breeding stock were excluded.

### 3.2.3 Puppy Raiser Questionnaire - C-BARQ

The Canine Behavioral Assessment and Research Questionnaire (C-BARQ) comprises 100 items which were grouped into seven sections as shown in Table 3.1. The questions were designed to allow participants to describe how their dog behaved in the recent past using a 5-point scale. Table B.2 shows each of the original questionnaire items and details of the rating scales.

Table 3.1: C-BARQ rating sections, items and number of items per section.

| Section | Items | Quantity |
|---|---|---|
| 1. Trainability | 1 - 8 | 8 |
| 2. Aggression | 9 - 35 | 27 |
| 3. Fear and Anxiety | 36 - 53 | 18 |
| 4. Separation-related behaviour | 54 - 61 | 8 |
| 5. Excitability | 62 - 67 | 6 |
| 6. Attachment and Attention-seeking | 68 - 73 | 6 |
| 7. Miscellaneous | 74 - 100 | 27 |
| **Total** | | **100** |

In order to retain information on the ordering of the original items, the data from the questionnaire were converted from categorical variables (factor data type) to discrete variables (numeric data type) from 0 to 4. Factors were derived by averaging the response to specific items; however, this calculation has evolved over time as Hsu et al. 2003 presented 11 factors and Duffy et al. 2012 introduced 14 factors.

Following the latest recommendations of the questionnaire's creators (Duffy et al. 2012), a total of 14 factors were calculated by averaging the original items as outlined in Table B.1. Out of the 100 original items presented in Table B.2, 78 (items 1–76 and 91–92) were used to calculate 14 factors, and the remaining 22 (items 77-90 and 93-100) were referred to as miscellaneous items (Duffy et al. 2012). All miscellaneous items (77-90 and 93-100) belong to the Miscellaneous section (74-100), but five items from the Miscellaneous section were used to calculate other factors. Specifically, factor "F_13 Chasing" depends on items 74, 75, and 76; and factor "F_14 Energy level" comprises items 91 and 92.

Two strategies were adopted to handle missing values in the calculation of the 14 factors described in Table B.1 similar to Duffy et al. 2012. Firstly, factors were only calculated if at least 75% of the items necessary for their calculation were available. For instance, Factor 1 would only be calculated for a particular dog if at least eight out of the ten essential items were available. Secondly, missing values were removed before taking their average for the calculation of the factors. In other words, the average formula took into account the number of items available for the calculation of the factor rather than the total number of items.

### 3.2.4 Dog Trainer Questionnaire - MCPQ-R

The Monash Canine Personality Questionnaire - Revised (MCPQ-R) is composed of 26 adjective-based items measuring canine personality along five dimensions: extraversion, motivation, training focus,

amicability, and neuroticism. The number of items per dimension is shown in Table 3.2. Table 3.3 shows the rating items that were rated by participants using a 6-point scale from 1 (really does not describe my dog) to 6 (really describes my dog).

Table 3.2: MCPQ-R rating dimensions, items, and the number of items per section.

| Dimensions | Items | Quantity |
|---|---|---|
| 1. Extraversion | 1 - 6 | 6 |
| 2. Motivation | 7 - 11 | 5 |
| 3. Training Focus | 12 - 17 | 6 |
| 4. Amicability | 18 - 22 | 5 |
| 5. Neuroticism | 23 - 26 | 4 |
| **Total** | | **26** |

Table 3.3: MCPQ-R rating items and items.

| Extraversion | Motivation | Training | Amicability | Neuroticism |
|---|---|---|---|---|
| Active | Assertive | Attentive | Easy going | Fearful |
| Energetic | Determined | Biddable | Friendly | Nervous |
| Excitable | Independent | Intelligent | Non-aggressive | Submissive |
| Hyperactive | Persevering | Obedient | Relaxed | Timid |
| Lively | Tenacious | Reliable | Sociable | |
| Restless | | Trainable | | |

### 3.2.5 Data Analysis

The data were analysed using R version 4.2.1 (Team 2022). Generalised Linear Models with a logit link function were used to implement the logistic regression models, for both feature selection and prediction of training outcome (Hosmer et al. 2013). They were referred to as logistic regression models for brevity henceforth. For the purpose of the analysis, the C-BARQ and MCPQ-R variables were processed, as numerical features representing interval variables as their scales were predetermined by design, varying from 0 to 4 and 1 to 6, respectively (Hsu et al. 2003; Ley et al. 2009b).

The original C-BARQ dataset comprised two demographic variables (sex and breed), 100 items and 14 factors, resulting in 116 features. The original MCPQ-R dataset contained two demographic variables and 26 items, resulting in a total of 28 features. The reduced feature sets were created from the original C-BARQ and MCPQ-R datasets, by selecting the features that met the following criteria:

1. a significant association ($p < 0.05$) with the training outcome as calculated by a logistic regression;
2. a significant reduction ($p < 0.05$) of deviance in the logistic regression model as evaluated by the chi-squared test;
3. items not included in the calculation of factors that met criteria 1 and 2 - C-BARQ only;
4. less than 5% missing values - C-BARQ only;
5. a correlation value below 0.75 with other independent variables.

The only features that achieved a significant level of association with training outcome were selected (i.e., criteria 1 and 2) as proposed in Asher et al. 2013. Unlike Duffy et al. 2012 and Bray et al. 2019, C-BARQ items that were not used in the calculation of significant factors were also included (criterion 3). This approach was employed for the following reasons: (1) it still prevented correlations between features, by removing original items used to calculate factors; (2) it allowed the use of original items whose derived factors were not significantly associated with training outcome; (3) it ensured the use of more general behaviour categories (factors) rather than the specific behaviours observed (items). Finally, variables containing more than 5% missing data were removed, instead of the 10% as utilised in Duffy et al. 2012 and Bray et al. 2019 due to the size of our dataset.

The logistic model becomes unstable if predictor variables are related, causing covariates to appear unimportant when all others are in the model (Aguilera et al. 2006). In order to remove highly multicollinear covariates, the absolute value of the pair-wise correlation between covariates was assessed. If the correlation reached the threshold of 0.75, the variable which had the highest mean absolute correlation considering all other covariates was removed.

A dimensionality reduction method is needed when the dataset considered for multivariate logistic regression has a high number of predictor variables relative to the number of observations. Principal Component Analysis (PCA) is a multivariate technique used to project the original data points onto an optimally calculated orthogonal coordinate system to capture the highest variance in the original dataset.

The reduced feature dataset was pre-processed in three steps as follows: imputation of missing values with the mean, normalisation of values using z-score, and finally dimensionality reduction with PCA. PCA was set to keep the principal components that captured 95% of the variance in the reduced set and discard the remainder. Finally, multivariate logistic regression was used to model the relationship between the independent variables in the reduced set, which was composed of $M$

features and $N$ examples and the dependent variable training outcome.

In order to maximise the number of training examples, the sampling strategy used was Leave One Out Cross Validation (LOOCV). The LOOCV algorithm fitted the model to the entire dataset excluding the example used for validation ($N-1$) until all $N$ examples were used for validation. The predictions made using the $N$ validation examples were then used to calculate the cross-validated performance metrics.

The model predicted the probability of withdrawal instead of success; hence, the case or positive class is the failure, and the control or negative class is the success. The confusion matrix for the predicted and true training outcome estimated by the model is shown in Figure 3.1. "True Fail" and "True Success" indicate the number of dogs that were correctly classified as a fail and success, respectively. On the other hand, "False Fail" and "False Success" denote the number of dogs that were incorrectly classified as a fail and success, respectively. "False Fail" encompasses dogs predicted to fail, but actually succeeded in their training. "False Success" considers dogs incorrectly classified as a success; however, they eventually failed their training.

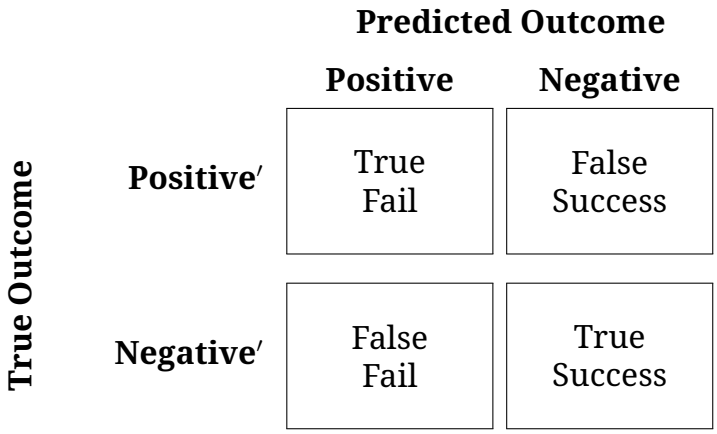



Figure 3.1: Confusion matrix for the predicted and true outcome.

The model was intended to aid assistance dog organisations when deciding whether to continue training a dog or release it from training by providing an objective metric (probability of withdrawal) at an early stage of the training. Hence, it was important to consider the supply of prospective dogs and demand for trained dogs, and the trade-off between the number of "False Fail" and "False Success" to determine which value was the most critical value for the application at hand (Jiang et al. 2008). "False Fail" causes a waste of the valuable resources already invested to get this prospective dog to this stage of the programme. "False Success" results in an organisation continuing to train a prospective dog that would eventually come to fail which caused

a waste of future resources until they do decide to withdraw the dog from training.

Accordingly, this model was designed to aid dog trainers by flagging only dogs that were very likely to be withdrawn; however, they were ultimately responsible for making a final decision. This is due to the fact that there was a high demand for trained dogs and a low supply of prospective dogs in the organisation involved in this study. Hence, predicting a fail outcome for a dog that would actually succeed ("False Fail") had more negative consequences than predicting success for a dog that would fail ("False Success"). Therefore, it was crucial to minimise the number of "False Fail" which indicates the number of dogs predicted as fail but actually succeeded in training.

Model performance was evaluated using the following metrics considering the values presented in the confusion matrix (Sokolova et al. 2009):

- Precision or positive predictive value (PPV) as shown in Equation 3.1. Precision is the proportion of dogs correctly classified as failures considering all the predicted fail examples. It denotes the class agreement between the predicted and true fail class.

$$Positive\ Predictive\ Value\ (PPV) = \frac{True\ Fail}{True\ Fail + False\ Fail} \tag{3.1}$$

- Sensitivity, recall, or true positive rate (TPR) as indicated in Equation 3.2. Sensitivity is the percentage of dogs classified as failures considering all the true fail examples. It estimates the ability to identify the fail class.

$$True\ Positive\ Rate\ (TPR) = \frac{True\ Fail}{True\ Fail + False\ Success} \tag{3.2}$$

- Specificity or true negative rate (TNR) as described in Equation 3.3. Specificity is the percentage of dogs classified as a success considering all the true success examples. It estimates the ability to identify the success class.

$$True\ Negative\ Rate\ (TNR) = \frac{True\ Success}{True\ Success + False\ Fail} \tag{3.3}$$

In order to minimise the critical value "False Fail", the model should be optimised for either precision or specificity. The choice between them depends on whether it is more important to reliably predict failure (choose precision) or identify success (choose specificity). In this particular application, the deployment of this model would benefit

assistance dog organisations more by providing a reliable prediction of failure. This would encourage practitioners to assess dogs flagged as likely to fail more closely in order to decide whether they should be withdrawn from training. Therefore, precision is a more appropriate metric to be optimised.

A backward feature selection algorithm was implemented using Recursive Feature Elimination (RFE) to maximise the precision. This RFE algorithm started by fitting a model using all principal components ($M$) to create a ranking for the features, then it removed the lowest ranked feature at each iteration until only one feature was left (i.e., $M, M-1, ..., 2, 1$). The model which achieved the highest precision was selected as the optimal model.

Model fitness was analysed using the likelihood ratio test for nested models, which follows a chi-squared distribution. The likelihood ratio test statistic was calculated based on the difference between the deviance of the null model and the proposed model in Equation 3.4. The degrees of freedom used for the chi-squared test corresponded to the reduction in the degrees of freedom from the null to the proposed model, as shown in Equation 3.5. The proposed model was considered a good fit if the p-value resulting from the chi-squared test performed using the critical value $\chi^2 = D_{Difference}, df = df_{Difference}$ was smaller than the significance level of 0.05 (Hosmer et al. 2013).

$$D_{Difference} = D_{Null} - D_{Proposed} \tag{3.4}$$

$$df_{Difference} = df_{Null} - df_{Proposed} \tag{3.5}$$

Precision, sensitivity, and specificity are one-dimensional or partial performance metrics since they only take into account the results from either the positive or the negative class, but not both (Luque et al. 2019). It is important to estimate the performance of the optimal model using a metric that takes into account this trade-off. In particular, the Precision-Recall (PR) curve considers the trade-off between precision and sensitivity, and the Receiver Operating Characteristic (ROC) curve is a measure of sensitivity and specificity (Jiang et al. 2008).

The areas under the PR and ROC curves provide a single measure reflecting the balance between metrics. Their values range between 0.5 and 1, with a value of 0.5 indicating a non-informative model, and a value of 1 indicating a perfectly predictive model. The area under the ROC analysis does not have any bias toward models that perform well on the minority class at the expense of the majority class which makes it quite attractive when dealing with imbalanced data (Luque et al. 2019). It has been reported as the most commonly used metric for imbalanced datasets (Haixiang et al. 2017).

Conclusively, precision was used to optimise the model, and the area under the ROC curve was used to validate and compare models. Sensitivity, specificity, and area under the PR curve were also reported. All metrics were calculated using the predictions on the validation set using LOOCV.

A flagging system similar to that presented in Harvey et al. 2017; Asher et al. 2013 and Bray et al. 2019 was utilised. Arbitrary thresholds were chosen to improve the predictive value of the negative and positive classes. The probability range was divided into three areas: dogs in the lower 0.30 range were flagged green, dogs in the middle 0.40 range were not flagged, and the top 0.30 range were flagged yellow. Therefore, dogs were assigned a yellow flag if their probability of withdrawal was above an arbitrary threshold of 0.70 and a green flag if their probability of withdrawal was below an arbitrary threshold of 0.30.

The impact that the features from the reduced dataset have on the principal components was investigated through the coefficient loadings on the principal components selected by the RFE. The estimates from the optimal multivariate logistic regression model were analysed through the calculation of the Odds Ratio (OR) and 95% Confidence Interval (CI). The interpretation of the optimal model was limited due to the data transformations applied; namely, normalisation and PCA.

## 3.3 Results

### 3.3.1 Puppy Raiser Questionnaire - C-BARQ

The puppy raisers of 63 dogs rated their dog's behaviour using C-BARQ. In this sample, the breakdown of the dogs' sex was female (N = 30) and male (N = 33); breed was Labrador Retriever (LR) (N = 25), Labrador Retriever crossed with Golden Retriever (LRxGR) (N = 36), Golden Doodle (N = 1), German Shepherd (N = 1); and age at assessment in months was (M = 16.1, SD = 3.4). The Golden Doodle and German Shepherd were combined into one class called "Other" for further analysis. The failure rate in this sample was 18%, as 11 failed and 52 succeeded in their training.

Figure 3.2 shows the duration of training for the dogs withdrawn from training for behavioural reasons. Dogs were withdrawn after 14 weeks of training on average. Only three out of the total 11 dogs that failed to graduate were withdrawn before 10 weeks.

Table B.3 shows the descriptive statistics including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing data of all original items and derived factors summarised by section. Although the scoring scale ranged between 0 and 4, most variables had

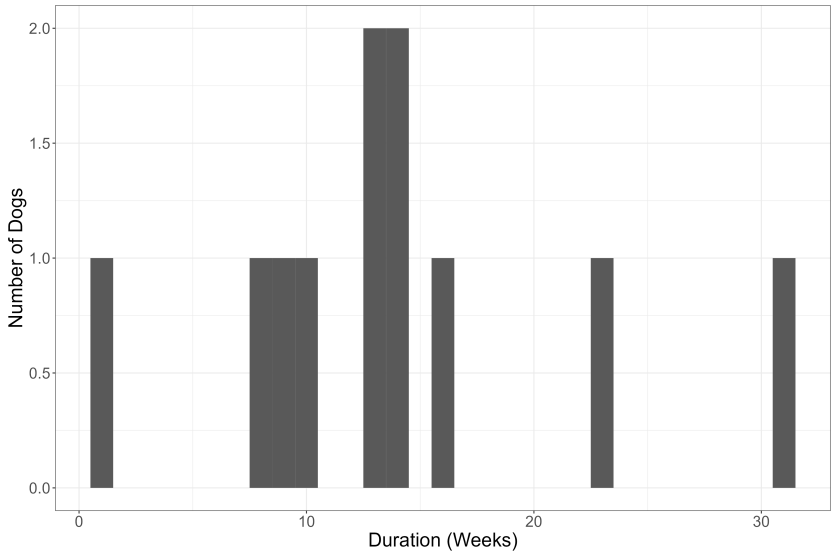



Figure 3.2: Duration of training in weeks for dogs that were withdrawn from training.

a much lower range of values recorded. Some variables were constant.

Some variables had a considerable amount of data missing due to the nature of the behaviours or situations analysed. Specifically, variables containing more than 10% missing values are shown in Table 3.4. The only factor with more than 5% data missing was "F_04 Dog rivalry", which was calculated based on items I_32, I_33, I_34, and I_35 as shown in Table B.1. Although NAs were expected, only in the case where there was no other household or familiar dogs, these items varied from 10 to 12 missing values. Therefore, five out of the 10 variables containing more than 10% missing data were due to not having another household dog.

#### 3.3.1.1 Feature Selection

Univariate logistic regression was used to investigate whether there was a statistically significant difference between the training outcome of the dogs and C-BARQ rating items and factors, and their demographic information. The results from the logistic regression univariate models are shown in Table B.4 where the 29 variables satisfying the first two criteria are highlighted.

Demographic information, such as breed (Other vs. LR p = 0.44 and Other vs LRxGR p = 0.16) and sex (p = 0.61), was not significantly associated with training outcome; hence, both variables were discarded (criterion 1). A total of 29 variables from C-BARQ met criteria 1 and 2 which indicated a significant association with training outcome and reduction in deviance, respectively. Specifically, there were 23 original items (out of 100) and six derived factors (out of 14).

Table 3.4: Variables with more than 5% missing values in the C-BARQ dataset.

| Variable | NAs (%) |
|---|---|
| I_49 When having nails clipped by a household member | 28 (44%) |
| I_76 Chases or would chase squirrels rabbits and other small animals given the opportunity | 18 (28%) |
| I_34 When approached while eating by another familiar household dog NA if no other dogs | 12 (19%) |
| I_33 When approached at a favourite resting or sleeping place by another familiar household dog NA if no other dogs | 11 (17%) |
| F_04 Dog rivalry | 11 (17%) |
| I_32 Towards another familiar dog in your household NA if no other dogs | 10 (16%) |
| I_35 When approached while playing with or chewing a favourite toy bone object etc by another familiar household dog NA if no other dogs | 10 (16%) |
| I_74 Chases or would chase cats given the opportunity | 7 (11%) |
| I_93 Stares intently at nothing visible | 7 (11%) |
| I_100 Displays other bizarre strange or repetitive behaviours | 7 (11%) |
| I_05 Slow to respond to correction or punishment thickskinned | 6 (10%) |
| I_09 When verbally corrected or punished scolded shouted at etc by you or a household member | 6 (10%) |
| I_77 Escapes or would escape from home or yard given the chance | 6 (10%) |
| I_96 Chases or follows shadows light spots etc | 5 (8%) |
| I_52 When unfamiliar dogs visit your home | 4 (6%) |
| I_94 Snaps at invisible flies | 4 (6%) |
| F_13 Chasing | 4 (6%) |

Subsequently, criterion 3 was applied to the 23 items that met criteria 1 and 2. It entailed keeping only items not included in the calculation of the six significant factors as indicated in Table B.1. In particular, items 1, 27 and 75 from the range of items used for calculating the factors (1-76 and 91-92) and items 77, 82, 85, 90, 93 and 99 from the range not used

for calculating factors (77-90, 93-100) were considered for the next stages. Hence, nine original items and six derived factors satisfied the first three criteria.

Criterion 4 was applied to the 15 variables meeting criteria 1, 2, and 3, which involved analysing missing data. The number of missing data for each of the variables is shown in Table 3.5. Considering that the original dataset contained 63 examples, two variables were removed as they contained more than 5% missing data (i.e., more than three examples); namely, "I_93 Stares intently at nothing visible" and "I_77 Escapes or would escape from home or yard given the chance".

Table 3.5: Variables considered for the reduced feature set and the number of missing values.

| Variables | NAs |
|---|---|
| I_01 When off the leash returns immediately when called | 0 |
| I_27 Toward cats squirrels or other small animals entering your yard | 2 |
| I_75 Chases or would chase birds given the opportunity | 1 |
| I_77 Escapes or would escape from home or yard given the chance | 6 |
| I_82 Begs persistently for food when people are eating | 1 |
| I_85 Pulls excessively hard when on the leash | 1 |
| I_90 Hyperactive restless has trouble settling down | 1 |
| I_93 Stares intently at nothing visible | 7 |
| I_99 Licks people or objects excessively | 3 |
| F_02 Stranger directed aggression | 0 |
| F_06 Nonsocial fear | 0 |
| F_07 Dog directed aggression | 0 |
| F_08 Dog directed fear | 0 |
| F_10 Separation related behaviour | 1 |
| F_11 Excitability | 0 |
| Outcome | 0 |

Finally, criterion 5 was applied by calculating the correlation between those 13 variables and removing highly correlated variables. The correlation heatmap was shown in Figure 3.3. Only one variable was removed for reaching a correlation level of more than 0.75, namely "F_07 Dog directed aggression", because of its high correlation with "F_08 Dog directed fear".

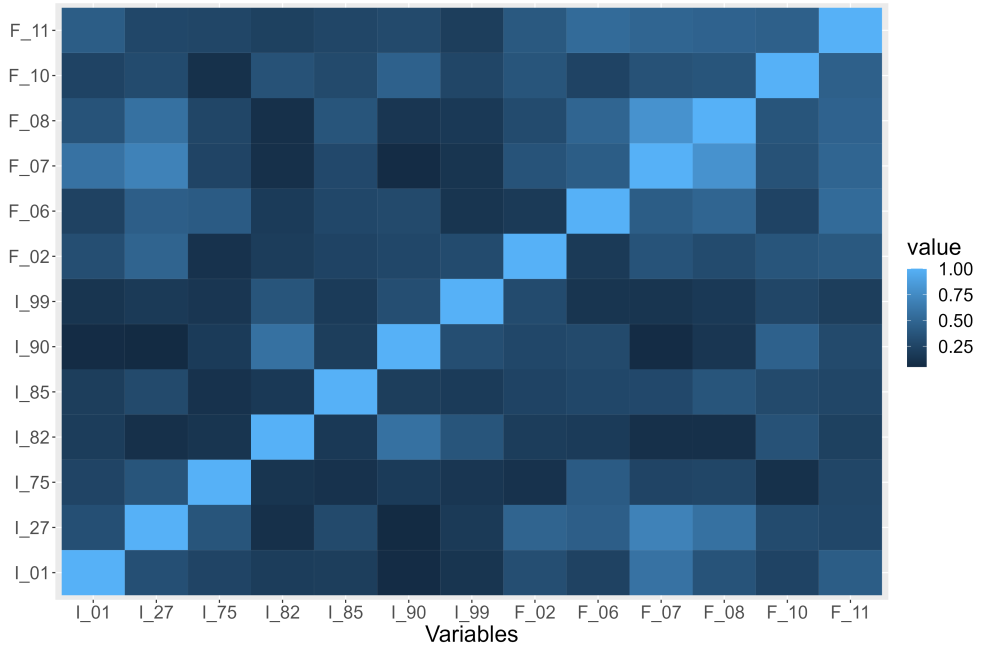



Figure 3.3: Correlation matrix of variables considered for the C-BARQ reduced feature set.

Consequently, the C-BARQ reduced feature set was created with 12 variables, more specifically seven items and five factors. Figures B.1 to B.12 show bar charts and violin box plots of each of the variables selected for the reduced feature set categorised by training outcome.

#### 3.3.1.2 Model

Data pre-processing techniques were applied to the reduced feature set, including normalisation, imputation, and PCA. PCA was implemented to capture 95% of the variance; therefore, the 10 first principal components were kept and the last two were removed. The explained variance per principal component is shown in Figure 3.4. PC01, the first principal component, captured 34% of variance in the data, which was higher than the next three largest components combined.

The optimal logistic regression model selected by RFE contained a single principal component (PC01) as a predictor. The parameters estimated for the optimal model using PC01 are shown in Table 3.6. The model estimates were exponentiated to get the OR and the 95% CI was also calculated. These values were interpreted to understand the relationship between PC01 and training outcomes.

The OR for PC01 is 0.45; this means that the probability of withdrawal decreases by 55% ($0.45 - 1 = -0.55$) for every unit increase of PC01. There is a 95% confidence that a one unit increase in PC01 results in 32 to 74% ($0.68 - 1 = -0.32$, $0.26 - 1 = -0.74$) lower odds of being withdrawn

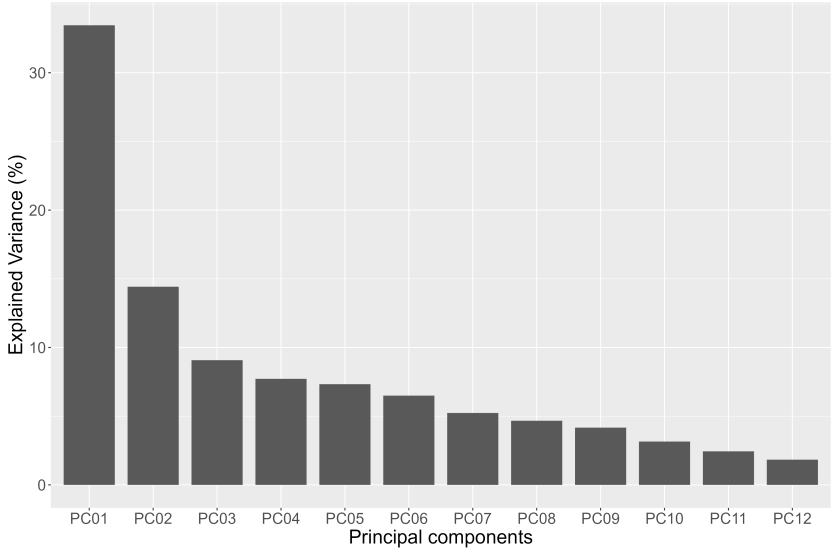



Figure 3.4: Explained variance per principal component (Scree Plot).

Table 3.6: Best logistic regression model using selected by RFE.

| | Estimate | Std. Error | z value | p-value | OR | 2.5% | 97.5% |
|---|---|---|---|---|---|---|---|
| (Intercept) | -2.06 | 0.47 | 4.40 | 0.00 | 0.13 | 0.04 | 0.29 |
| PC01 | -0.80 | 0.24 | 3.31 | 0.00 | 0.45 | 0.26 | 0.68 |

from training.

Model fitness was evaluated considering the reduction in deviance resulting from the addition of one variable (PC01). The null deviance was 58.35 on 62 degrees of freedom, and the residual deviance was 39.33 on 61 degrees of freedom. Hence, the difference in deviance was $58.35 - 39.33 = 19.02$ on $62 - 61 = 1$ degrees of freedom. The model can be considered a good fit for the data as it fitted the data significantly better than the null model $P(\tilde{\chi}_1^2 \geq 19.02) = 1.30e^{-05}$. Thus, this significant reduction in deviance supported the hypothesis that PC01 was a significant variable in predicting training outcomes.

#### 3.3.1.3 Performance

The ROC and PR curves are shown in Figure 3.5. The area under the ROC curve was 0.84 and the PR curve was 0.64. The precision, specificity, and sensitivity estimated using LOOCV prediction results were 0.57, 0.94, and 0.36, respectively.

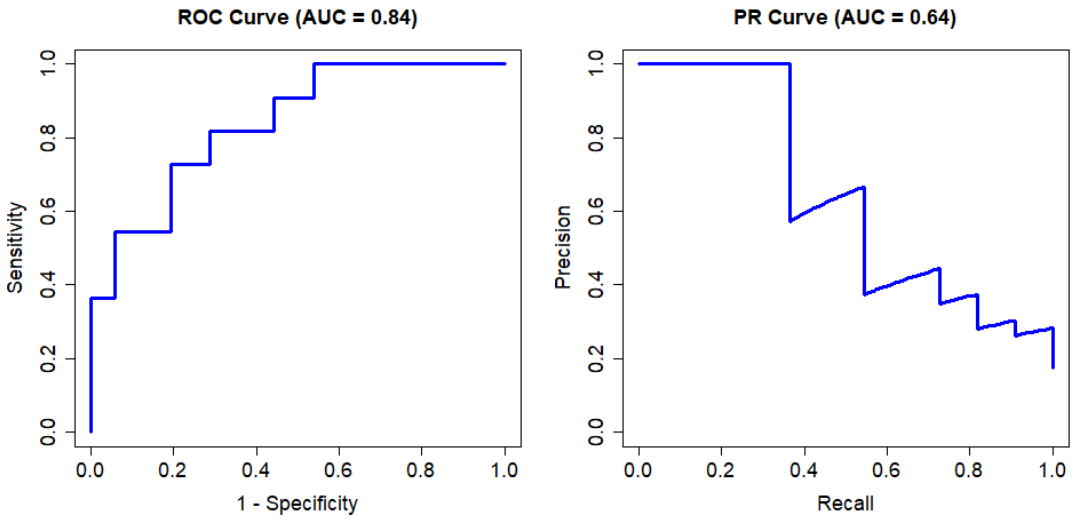



Figure 3.5: ROC and PR curves for the best logistic regression model using C-BARQ data selected by RFE.

#### 3.3.1.4 Application

Figures 3.6 and 3.7 show the probability of failure predicted by the best logistic regression model for all the 63 dogs in the dataset considering the 12 C-BARQ items and factors as rated by their puppy raisers. Two boxes show idealised thresholds for a flagging system, where the green flag is assigned to dogs likely to succeed and the yellow flag to dogs likely to be withdrawn. The arbitrary thresholds selected were 0.30 for training success and 0.70 for withdrawal as described in Section 3.2.5.

Figure 3.6 displays the training outcome obtained by each dog, i.e., whether the dog failed or succeeded in their training. It demonstrates that the lowest estimates belonged to dogs who succeed and the highest estimates to dogs that failed. Considering the flagging system in the cross-validated results, green flags were assigned to 53 dogs and yellow flags to four dogs. From the green-flagged group, only five dogs failed their training resulting in 91% precision in predicting success. In the yellow-flagged group, all four dogs failed, yielding 100% precision in the prediction of failure.

Figure 3.7 shows the same probability curve highlighting the status of training outcome, i.e., whether the dog became a guide dog, service dog, or was withdrawn from training. The remaining six dogs that scored between 0.30 and 0.70 got a very mixed outcome as expected; two were withdrawn, two became assistance dogs, and two became guide dogs. Guide dogs achieved a lower estimated probability of withdrawal than service dogs. In particular, guide dogs had an average probability of withdrawal of 0.10, service dogs 0.14, and withdrawn dogs 0.44. It can

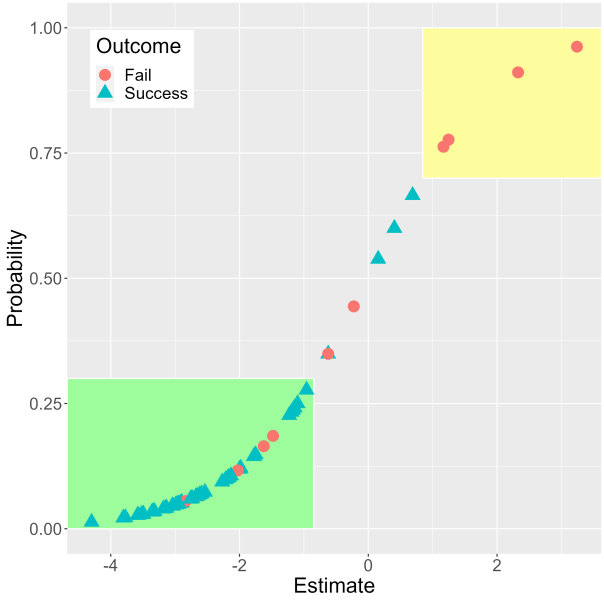



Figure 3.6: Probability of withdrawal predicted by the best logistic regression model using C-BARQ data per true outcome (fail, success).

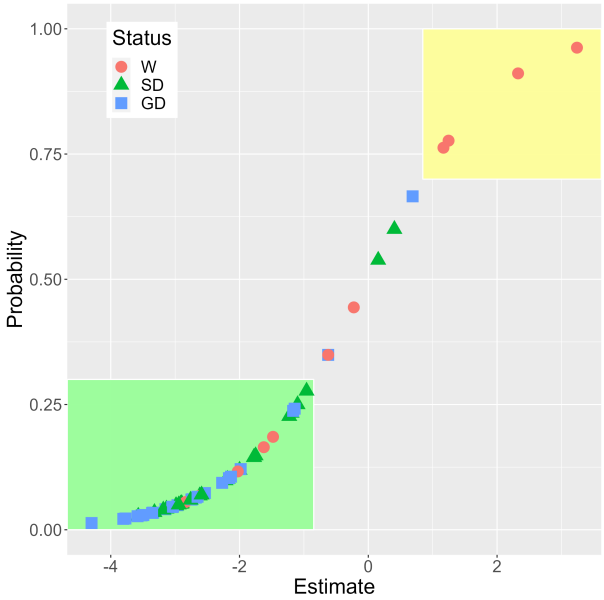



Figure 3.7: Probability of withdrawal predicted by the best logistic regression model using C-BARQ data per true status (withdrawn, service dog or guide dog).

be visually observed that non-flagged dogs were a mix of guide, service, and withdrawn dogs.

#### 3.3.1.5 Feature Importance

Principal components can be interpreted according to their coefficient loadings as they represent the level of correlation between the principal

component and the original variables. The coefficient loading on PC01 is shown in Figure 3.8. The only positive coefficient was related to the variable "I_01 When off the leash returns immediately when called" while all others were negative. This result indicates that higher scores in I_01 produce higher values of PC01; however, higher scores in any of the other original variables cause lower values of PC01.

The top five variables with the highest coefficient loading are "F_11 Excitability", "F_08 Dog directed fear", "F_06 Nonsocial fear", "F_10 Separation related behaviour", and aggression "I_27 Toward cats squirrels or other small animals entering your yard". By definition, the higher the coefficient loading the stronger the correlation between that original variable and the principal component. Hence, these are the variables that most strongly contribute to PC01.

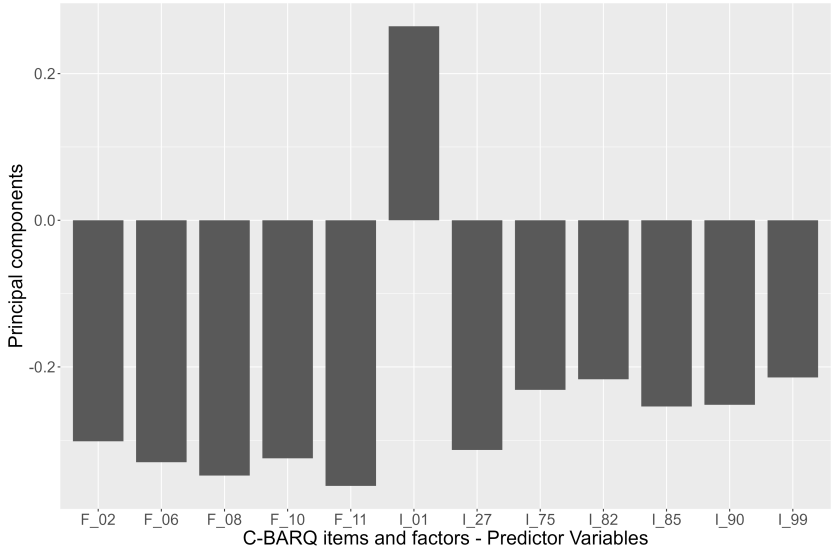



Figure 3.8: PC01 coefficient loading from original items and factors.

### 3.3.2 Dog Trainer Questionnaire - MCPQ-R

The dog trainers of 89 dogs rated their dog's behaviour using MCPQ-R. In this sample, the breakdown of the dogs' sex was female (N = 44) and male (N = 45); breed was Labrador Retriever (LR) (N = 37), Labrador Retriever crossed with Golden Retriever (LRxGR) (N = 49), Golden Doodle (N = 2), and German Shepherd (N = 1); and age at assessment in months was (M = 18.34, SD = 2.6). A total of 71 dogs succeeded (70%) and 18 dogs failed (20%).

Figure 3.9 shows the duration of training for the dogs withdrawn from training for behavioural reasons. Dogs were withdrawn after 16 weeks of training on average, only three out of the 18 dogs that failed to graduate as service dogs were withdrawn before 10 weeks.

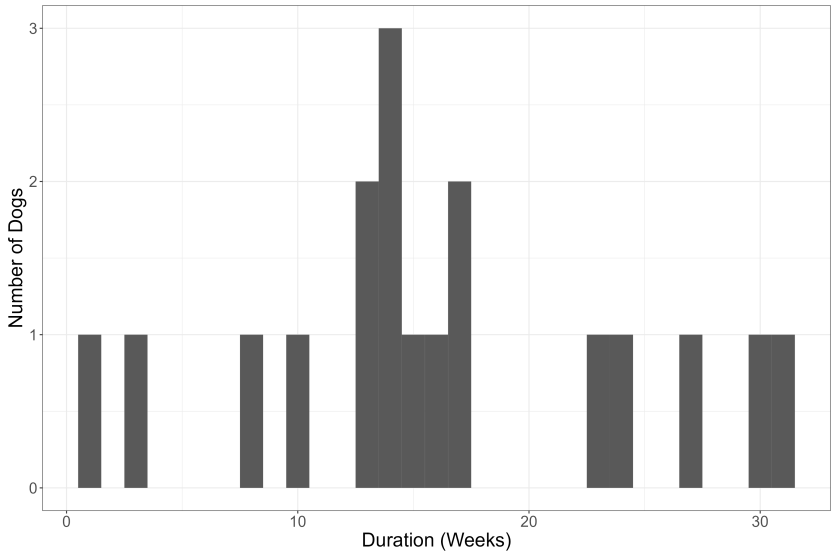



Figure 3.9: Duration of training in weeks for dogs that were withdrawn from training.

Table B.6 shows the descriptive statistics including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing data of all original items and derived factors summarised by section. Except for seven variables, all others recorded values in the entire range between 1 and 6. Due to the nature of the items in this questionnaire, there were no missing data.

#### 3.3.2.1 Feature Selection

Univariate logistic regression was used to investigate whether there was a statistically significant difference between the training outcome of the dogs and items and factors from the MCPQ-R rating and their demographic information. The results are shown in Table B.7 where the 13 variables satisfying the first two criteria are highlighted.

Demographic information, such as breed (Other vs. LR p-value = 0.82 and Other vs. LRxGR p-value=0.40) and sex (p-value = 0.32), was not significantly associated with training outcome. Criterion 1 and 2 were applied preserving only a total of 13 MCPQ-R variables from sub-scales Motivation, Training, Amicability, and Neuroticism; no items from Extraversion showed a significant relationship with training outcome.

The correlation between the variables was evaluated and is shown in Figure 3.10. Criterion 5 was applied resulting in the removal of one highly correlated item, namely “Neuroticism Fearful” due to its high correlation with “Neuroticism Nervous”.

Consequently, the MCPQ-R reduced feature set was created with 12 variables. Figures B.13 to B.24 show bar charts and violin box plots of

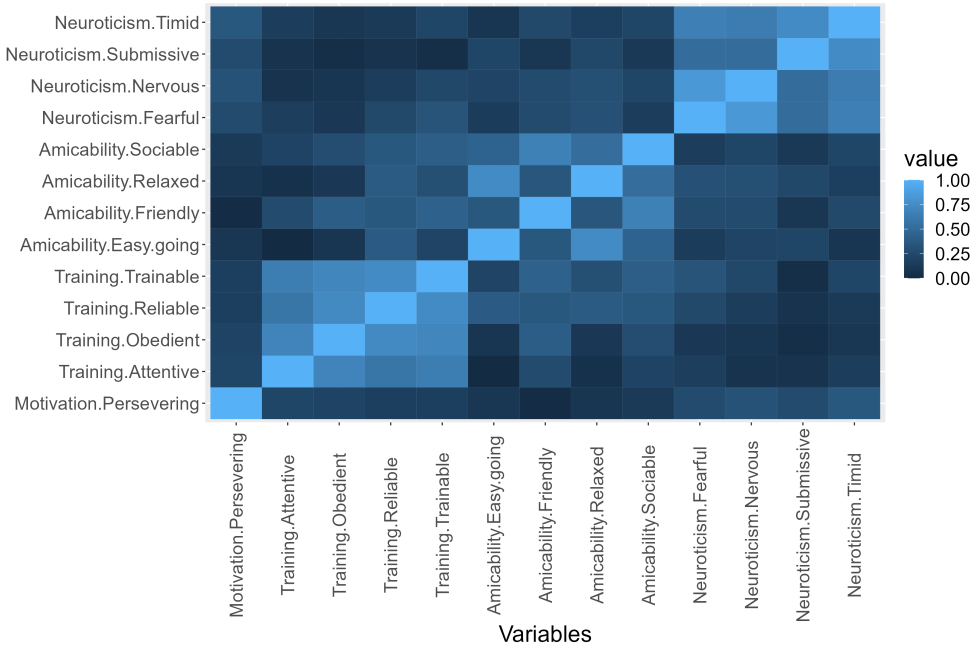



Figure 3.10: Correlation Matrix for variables considered for the MCPQ-R reduced feature set.

each of the variables selected for the MCPQ-R reduced feature set categorised by training outcome.

#### 3.3.2.2 Model

Data pre-processing techniques were applied to the reduced feature set, including normalisation, imputation, and PCA. PCA was implemented to capture 95% of the variance; consequently, it selected the first nine principal components and removed the last three components. The explained variance per principal component is shown in Figure 3.11. PC1, the first principal component, captured 36% of variance in the data, which was higher than the next two largest components combined.

The optimal logistic regression model selected by RFE contained a single principal component (PC1) as a predictor. The parameters estimated for the optimal model using PC01 are shown in Table 3.7. The model estimates were exponentiated to get the OR and the 95% CI was also calculated. These values are interpreted to understand the relationship between PC1 and training outcomes.

The OR for PC1 is 2.17; this means that the probability of withdrawal increases by ($2.17 - 1 = 1.17$) 117% for every unit increase of PC1. There is a 95% confidence that a one unit increase in PC1 results in a 54 to 237% ($1.54 - 1 = 0.54$, $3.37 - 1 = 2.37$) higher odds of being withdrawn from training.

Model fitness was evaluated considering the difference in deviance

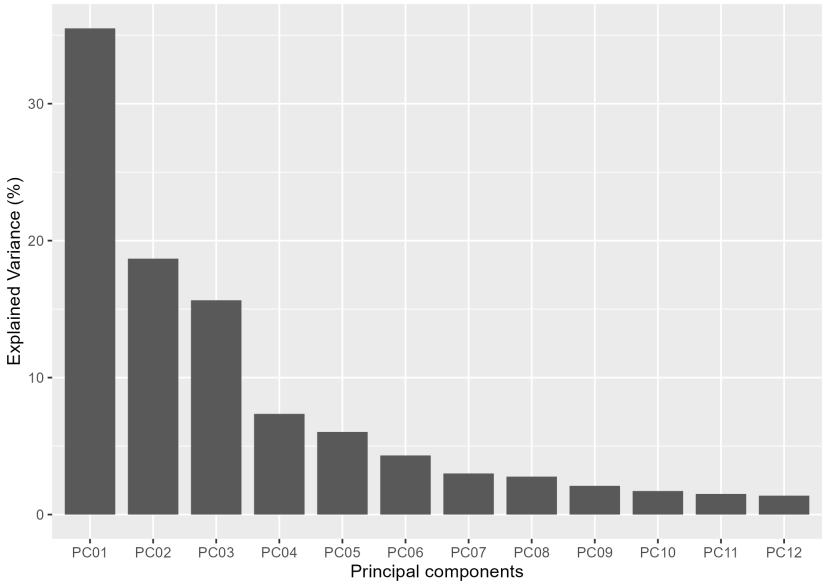



Figure 3.11: Explained variance per principal component (Scree Plot).

Table 3.7: Best logistic regression model using selected by RFE.

| | Estimate | Std. Error | z value | p-value | OR | 2.5% | 97.5% |
|---|---|---|---|---|---|---|---|
| Intercept | -1.99 | 0.40 | -4.97 | 0.00 | 0.14 | 0.06 | 0.28 |
| PC1 | 0.77 | 0.20 | 3.95 | 0.00 | 2.17 | 1.54 | 3.37 |

resulting from the addition of one variable (PC1). The null deviance was 89.62 on 88 degrees of freedom and the residual deviance was 63.58 on 87 degrees of freedom. Thus, the difference in deviance was $89.62 - 63.58 = 26.04$ on $88 - 87 = 1$ degrees of freedom. The model can be considered a good fit for the data as it fitted the data significantly better than the null model $P(\tilde{\chi}_2^2 \geq 26.04) = 3.34e - 07$. Thus, this significant reduction in deviance supported the hypothesis that PC1 was a significant variable in predicting training outcomes.

#### 3.3.2.3 Performance

The ROC and PR curves are shown in Figure 3.12. The area under the ROC curve was 0.85 and the PR curve was 0.58. The precision, specificity, and sensitivity estimated using LOOCV results were 0.75, 0.97, and 0.33, respectively.

#### 3.3.2.4 Application

Figures 3.13 and 3.14 show the probability of failure predicted by the best logistic regression model for all the 89 dogs in the dataset considering the 12 MCPQ-R items as rated by their dog trainers. Two

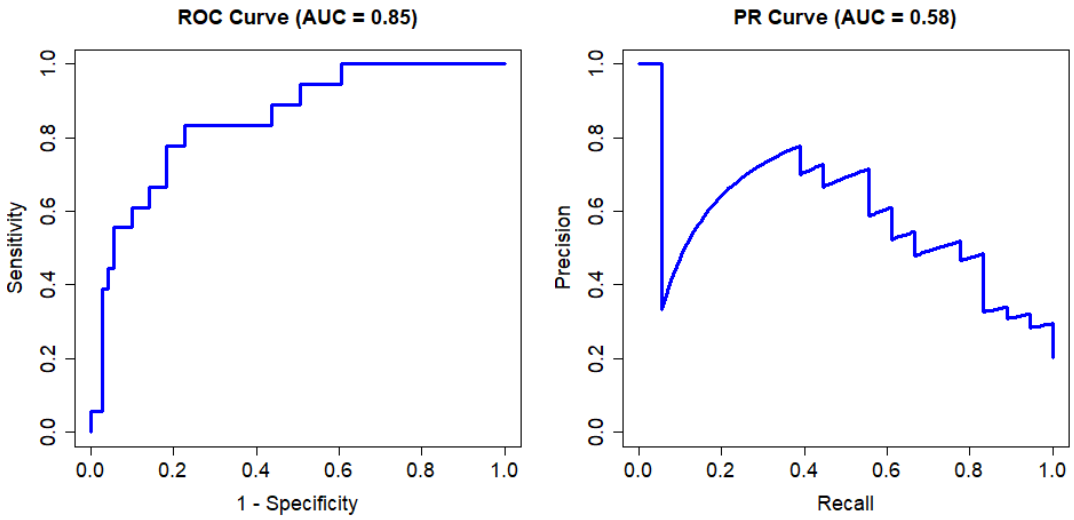



Figure 3.12: ROC and PR curves for the best logistic regression model using MCPQ-R data selected by RFE.

boxes show idealised thresholds for a flagging system, where the green flag was assigned to dogs likely to succeed and the yellow flag indicates that dogs were highly likely to be withdrawn. The thresholds selected were 0.30 for training success and 0.70 for withdrawal.

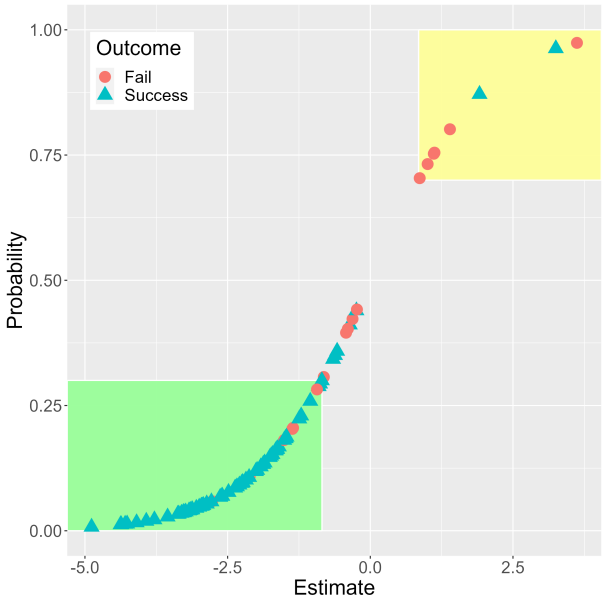



Figure 3.13: Probability of withdrawal predicted by the best logistic regression model using MCPQ-R data per true outcome (fail, success).

Figure 3.13 displays the training outcome obtained by each dog, i.e., whether the dog failed or succeeded in their training. It demonstrates that the lowest estimates belonged to dogs who succeed and the highest

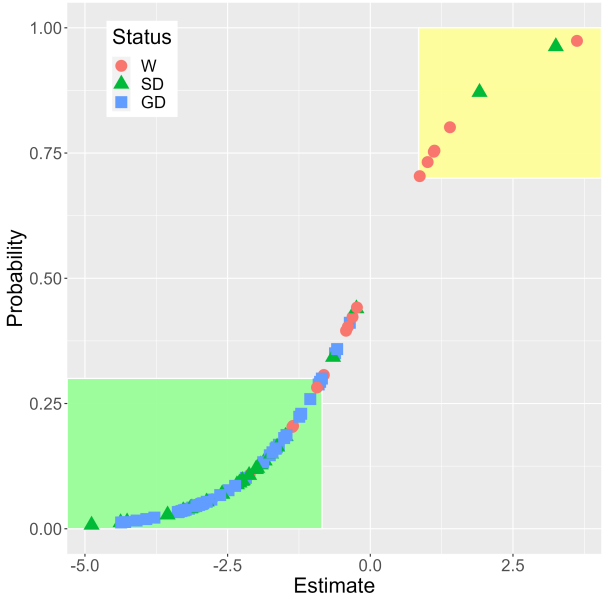



Figure 3.14: Probability of withdrawal predicted by the best logistic regression model using MCPQ-R data per true status (withdrawn, service dog or guide dog).

estimate to dogs that failed. Considering the flagging system in this idealised demonstration, green flags were assigned to 71 dogs and yellow flags to six dogs. From the green-flagged group, a total of 64 dogs succeeded and only seven failed their training resulting in 90% precision in predicting success. In the yellow-flagged group, six dogs failed and two dogs succeeded yielding 75% precision in the prediction of failure.

Figure 3.14 shows the same probability curve, but instead, it highlights the status of training outcome, i.e., whether the dog became a guide dog, service dog, or was withdrawn from training. The remaining 10 dogs scoring between 0.30 and 0.70 got a mixed outcome as expected; five were withdrawn from training, two graduated as service dogs, and two graduated as guide dogs. Guide dogs achieved a lower estimated probability of withdrawal than service dogs. In particular, guide dogs had an average estimated probability of withdrawal of 0.13, service dogs 0.17 and withdrawn dogs 0.44. It can be visually observed that non-flagged dogs were all service dogs.

#### 3.3.2.5 Feature Importance

Principal components can be interpreted according to their coefficient loadings as they represent the level of correlation between the principal component and the original variables. The coefficient loading on PC1 as calculated by PCA is shown in Figure 3.15. The only positive coefficients were related to variables in the sub-scale Neuroticism, all others were negative. Therefore, higher scores in any of the items in the Neuroticism

sub-scale; namely, Nervous, Sensitive, and Timid, produce higher values of PC1; however, higher scores in any of the other original variables will cause lower values of PC1.

The top 5 variables with the highest coefficients are related to the Training Focus (Trainable, Reliable, Obedient), and Amicability (Sociable, Friendly) sub-scales. By definition, the higher the coefficient loading the stronger the correlation between that original variable and the principal component. Hence, these are the variables that most strongly contribute to PC1.

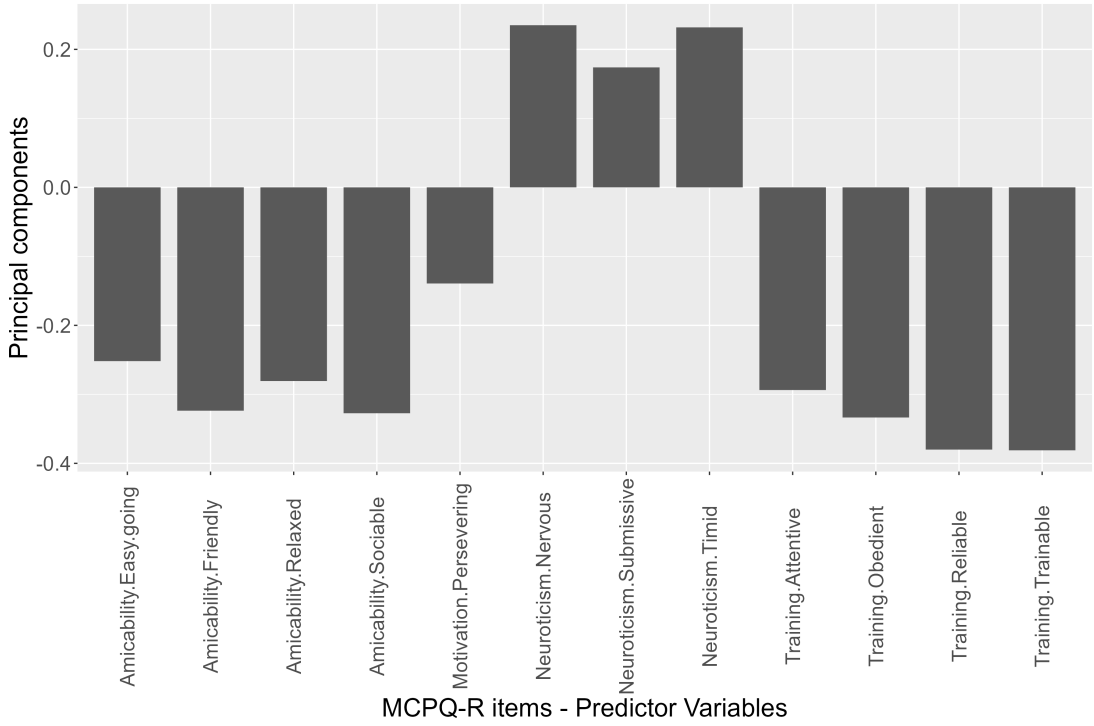



Figure 3.15: PC1 coefficient loading from original items.

## 3.4 Discussion

This chapter investigated the predictive power of two standardised behavioural ratings of prospective assistance dogs to estimate their training outcomes. This study demonstrated the benefit of collecting data on trainee assistance dogs using different canine behaviour and personality questionnaires as it allows the creation of statistical models for the prediction of training outcomes. Puppy raisers answered the C-BARQ around the time when dogs started their formal training (16 months) and dog trainers answered the MCPQ-R after 10 weeks of training (18 months). Machine learning algorithms were independently developed to predict training outcomes based on the data from each of the ratings separately.

This study demonstrated for the first time that the MCPQ-R can be used for estimating a dog's probability of withdrawal from training with comparable performance to the well-established C-BARQ. Moreover, the C-BARQ model significantly outperformed all models previously presented in the literature. This advancement is attributed to the application of various machine learning methods (feature selection and reduction techniques) in this field of research for the first time. This is the first study to validate the use of the MCPQ-R for the prediction of training outcomes in trainee assistance dogs. This finding allows dog training organisations to get an early prediction of training outcomes, through not only the puppy raiser's knowledge but also that of the dog trainer.

The models built on C-BARQ ratings completed by puppy raisers and MCPQ-R completed by dog trainers can greatly help assistance dog training organisations, by classifying dogs into green and yellow flagged groups before they start formal training, and at 10 weeks of training, respectively. The high predictive value of the yellow and green flags indicated that yellow-flagged dogs were very likely to fail, and green-flagged dogs were very likely to succeed. This allows a valuable practical application in this industry allowing for resources to be more optimally managed. In particular, dogs assigned a yellow flag can be more closely assessed, and problematic behaviours can be intently addressed at an early stage of training. Whenever training resources are limited (e.g., personnel, material, financial), the training of dogs assigned a green flag can be prioritised.

The methodology and results obtained from the C-BARQ and MCPQ-R are discussed in Sections 3.4.1 and 3.4.2. The relationship between each of the rating's original variables and the training outcome predictions are complex due to the data reduction technique employed (PCA). However, a high-level qualitative interpretation is possible considering the association between original variables and principal components as described in Figures 3.8 and 3.15 and the relationship between the principal components and probabilities estimated by the logistic regression model as shown in Tables 3.6 and 3.7.

The results obtained from the C-BARQ and MCPQ-R models developed in this research are compared in Section 3.4.3. A comprehensive comparison with previous research is provided in Section 3.4.4. It examines the model created in this research with those from previous research on training outcome prediction using the C-BARQ rating. Then, the prediction performance of both C-BARQ and MCPQ-R models designed in this research is compared with previous research using diverse questionnaires for the prediction of training outcomes in prospective assistance dogs.

### 3.4.1 Puppy Raiser Questionnaire - C-BARQ

The original C-BARQ dataset contained 100 items, 14 factors and two demographic variables. All demographic variables (sex and breed) and C-BARQ factors showing a significant relationship with the outcome ($p < 0.05$), as described in Table B.4 were considered for the reduced dataset. Only original items significantly related to outcome ($p < 0.05$) that were not used for the calculation of any significant factor were selected for the reduced dataset. Variables with more than 5% missing values or 75% correlation with other variables were removed. Consequently, the reduced C-BARQ dataset contained 12 features, specifically they were seven items and five factors.

The relationship between these variables and training outcome was assessed based on the sign (positive or negative) of the estimate in the univariate logistic regression models shown in Table B.4. A positive sign indicates a positive relationship (higher scores predicted a higher probability of withdrawal), and a negative sign expresses an inverse association (higher scores predicted a lower probability of withdrawal).

In particular, "I_01 When off the leash returns immediately when called" had a negative estimate indicating a reverse relationship between its score and probability of withdrawal (i.e., higher scores predicted lower probability of withdrawal). All the other 11 variables; namely, "I_27 Toward cats squirrels or other small animals entering your yard", "I_75 Chases or would chase birds given the opportunity", "I_82 Begs persistently for food when people are eating", "I_85 Pulls excessively hard when on the leash", "I_90 Hyperactive restless has trouble settling down", "I_99 Licks people or objects excessively", "F_02 Stranger directed aggression", "F_06 Nonsocial fear", "F_08 Dog directed fear", "F_10 Separation related problems", and "F_11 Excitability" had a positive estimate, denoting a positive relationship between the score and probability of withdrawal (i.e., higher scores predicted a higher probability of withdrawal).

Alternatively, this relationship can be visually assessed in Figures B.1 to B.12. The only variable for which withdrawn dogs scored lower than successful dogs was "I_01". Withdrawn dogs were rated higher than successful dogs for all other 11 variables in the reduced feature set. The results from the PCA indicate that the only variable with a positive coefficient loading on PC01 was "I_01" while all the other 11 variables had negative coefficient loadings as shown in Figure 3.8. The top five variables with the highest coefficients on PC01 were "F_11 Excitability", "F_08 Dog directed fear", "F_06 Nonsocial fear", "F_10 Separation related behaviour", and aggression "I_27 Toward cats squirrels or other small animals entering your yard". Effectively, these were the variables that most strongly influence the value of PC01 and consequently the probability of withdrawal.

PC01 is then multiplied by a negative logistic regression estimate (-0.80) as described in Table 3.6, therefore positive values of PC01 result in negative point estimates. The more negative a point estimate, the smaller the probability of withdrawal as indicated in Figure 3.6. An increase in PC01 decreases the probability of withdrawal while a decrease in PC01 increases the probability of withdrawal.

In summary, this indicates that the optimal logistic regression model learnt that the lower the scores in "I_01", or the higher the scores in any of the other 11 variables, the more likely the dog is to be withdrawn from their training programme. Hence, "I_01" is associated with PC01 and training outcome in the opposite direction to all other variables.

### 3.4.2 Dog Trainer Questionnaire - MCPQ-R

The original MCPQ-R dataset comprised 26 items and two demographic variables. All variables that were significantly related to training outcome were considered for the reduced dataset (Criterion 1 and 2). The correlation between the variables shown in Figure 3.10 indicates that items from the same sub-scale have a higher level of correlation. Highly correlated variables (>75%) were removed (Criterion 5). Accordingly, the reduced MCPQ-R dataset contained 12 variables relating to four out of the five sub-scales.

The relationship between those variables with training outcome can be assessed based on the sign (positive or negative) of the estimate from the univariate logistic regression models shown in Table B.7. A positive sign indicates a positive relationship (higher scores predicted a higher probability of withdrawal) and a negative sign expresses an inverse association (higher scores predicted a lower probability of withdrawal). This relationship can also be visualised in Figures B.13 to B.24.

Specifically, withdrawn dogs scored lower than successful dogs in all items in the sub-scales Motivation (Persevering), Training Focus (Attentive, Obedient, Reliable, Trainable), and Amicability (Easy going, Friendly, Relaxed, Sociable). Naturally, their estimates were negative indicating an opposite relationship between the scores and the probability of withdrawal (higher scores predicted a lower probability of withdrawal). On the other hand, withdrawn dogs scored higher than successful dogs in the Neuroticism sub-scale (Nervous, Submissive, Timid). Consequently, their estimates were positive denoting an analogous relationship between the scores and the probability of withdrawal (higher scores predicted a higher probability of withdrawal). No items from the Extraversion sub-scale were found to have a significant relationship with training outcomes.

The results from the PCA indicate that coefficient loadings on PC1 for all variables in the sub-scales Motivation, Training Focus, and Amicability

were negative, while positive coefficient loadings were associated with variables from the sub-scale Neuroticism, as shown in Figure 3.15. In particular, the top 5 MCPQ-R items with the biggest impact on PC1 and, consequently, on training outcome estimations are related to the Training Focus (Trainable, Reliable, Obedient) and Amicability (Sociable, Friendly). Effectively, these are the variables that most strongly influence the value of PC1 and consequently the probability of withdrawal.

The optimal logistic regression model estimate associated with PC1 is positive (0.77) as described in Table 3.7. Therefore, an increase in PC1 results in a higher probability of withdrawal and a decrease in PC1 results in a lower probability of withdrawal.

Conclusively, the optimal logistic regression model learnt that dogs scoring high in items related to the Neuroticism sub-scale and low in items related to Motivation, Training Focus, and Amicability sub-scales have a higher chance of being withdrawn from their training programme. In other words, variables belonging to the Neuroticism sub-scale were associated with PC1 and training outcomes in the opposite direction to all other variables.

### 3.4.3 Comparison between C-BARQ and MCPQ-R Models

This section compares the methodology utilised and the results achieved in the creation of the C-BARQ and MCPQ-R models. This was the first study to not only investigate the relationship between MCPQ-R and training outcomes but also develop a machine learning model to predict training outcomes using MCPQ-R data. This research demonstrated that the novel MCPQ-R model achieved similar results to the widely used C-BARQ.

From a practical perspective, collecting data for the C-BARQ is more challenging than MCPQ-R, due to the lower completion rates associated with longer questionnaires. As expected, the completion rate was lower for the C-BARQ filled out by the puppy raisers (46%) when compared to the MCPQ-R filled out by the dog trainers (74%). Two main reasons are attributed to this outcome: firstly, the length of the questionnaire as the MCPQ-R is significantly shorter (26 vs. 100 items); and secondly, the role of participants, as dog trainers were more involved with the training organisation than puppy raisers.

The items in the C-BARQ rating relate to the frequency or intensity of behaviours displayed in very specific scenarios. The items in the MCPQ-R rating relate to personality traits that characterise patterns of behaviour that are observed over time (Wiener et al. 2016). Hence, it can be expected that C-BARQ items need to be combined into more general personality traits, to distinguish between different training

outcomes. This assumption was confirmed by our results as a larger proportion of the C-BARQ calculated factors (6/14) showed statistical significance compared to the original items (23/100).

After applying the feature selection criteria, 12 out of 114 C-BARQ variables and 12 out of the 26 MCPQ-R items were kept for the dimensionality reduction and machine learning steps. The selected C-BARQ variables were seven items belonging to the following sections: Trainability ("I_01 When off the leash returns immediately when called"), Aggression ("I_27 Toward cats squirrels or other small animals entering your yard"), Miscellaneous ("I_75 Chases or would chase birds given the opportunity", "I_82 Begs persistently for food when people are eating", "I_85 Pulls excessively hard when on the leash", "I_90 Hyperactive restless has trouble settling down", "I_99 Licks people or objects excessively"); and five factors "F_02 Stranger directed aggression", "F_06 Nonsocial fear", "F_08 Dog directed fear", "F_10 Separation related problems", and "F_11 Excitability". The selected MCPQ-R variables were 12 items belonging to the following sub-scales: Motivation (Persevering), Training Focus (Attentive, Obedient, Reliable, Trainable), Amicability (Easy going, Friendly, Relaxed, Sociable), and Neuroticism sub-scale (Nervous, Submissive, Timid).

The similar performance results obtained by the MCPQ-R compared to the widely used C-BARQ provide a scientific basis for the use of the MCPQ-R for predicting training outcomes in assistance dogs. The C-BARQ model achieved an area under the ROC curve of 0.84 and PR curve of 0.64, a precision of 0.57, a specificity of 0.94, and a sensitivity of 0.36. The MCPQ-R model achieved an area under the ROC of 0.85 and PR curve of 0.58, a precision of 0.75, a specificity of 0.97, and a sensitivity of 0.33.

The proposed flagging system with two probability thresholds (lower=0.30 and upper=0.70) produced more precise predictions compared to using only one probability threshold of 0.50. In particular, the predictive value for withdrawal increased from 85% to 100% using the yellow flag for C-BARQ and remained 75% for MCPQ-R. The predictive value for success increased from 88% to 91% using the green flag for C-BARQ and from 85% to 90% for MCPQ-R.

The use of MCPQ-R rating offers several advantages over C-BARQ for the prediction of training outcomes in terms of data collection and modelling complexity. MCPQ-R questionnaire data is easier to obtain from dog trainers than C-BARQ from puppy raisers. A smaller number of raters are involved in the data collection, removing potential inter-rater variance in the data and simplifying the communication process. MCPQ-R questionnaire data is simpler to model because less pre-processing is required as it comprises fewer variables, contains no missing data and no factors need to be extracted. The only drawback is

that MCPQ-R data were collected at week 10 of training, while the C-BARQ data was collected before the dogs started training. However, this is a research design limitation which can be overcome in future research by following the recommendations presented in Section 6.3.

### 3.4.4 Comparison with Previous Research

This is the first study to examine the relationship between the MCPQ-R and training outcomes in trainee assistance dogs. Considering that the C-BARQ validated and widely used for the prediction of training outcomes (Duffy et al. 2008; Duffy et al. 2012; Bray et al. 2019; Hunt et al. 2020), machine learning models were separately developed using data from the C-BARQ and MCPQ-R. Comparing the results from the MCPQ-R and the C-BARQ on the same cohort of dogs allowed a fair comparison of the predictive power of these canine behaviour and personality assessment methods. The results of this study demonstrated for the first time that the MCPQ-R model achieved comparable prediction performance to the C-BARQ for estimating training outcomes in trainee assistance dogs. Accordingly, this research validated a novel application of the MCPQ-R for the prediction of training outcomes.

Additionally, this C-BARQ model surpassed the prediction performance of all models presented in previous research. The machine learning models were developed using various techniques for feature selection and dimensionality reduction for the first time in this field. Hence, this research improved the state-of-the-art performance in the prediction of training outcomes in trainee assistance dogs, by applying new data pre-processing techniques.

In this section, previous studies on the prediction of training outcomes in assistance dogs are compared to this study in terms of the relationship between independent variables and training outcomes, and model performance. In particular, six studies investigating the connection between C-BARQ and training outcomes are discussed in Section 3.4.4.1. The performance of different models for the prediction of training outcomes in assistance dogs built on data from various questionnaires is discussed in Section 3.4.4.2.

#### 3.4.4.1 C-BARQ and Training outcome

Six studies explored the relationship between C-BARQ and training outcomes in assistance dogs. In particular, two aimed at evaluating the association between the C-BARQ rating and training outcome (Duffy et al. 2008; Hunt et al. 2020), and four studies used the rating to predict training outcome (L. S. Batt et al. 2009; Duffy et al. 2012; Cleghern et al. 2018; Bray et al. 2019). However, it is not clear whether an overlap exists

in the datasets used in previous research, i.e., the datasets used in Duffy et al. 2008; Duffy et al. 2012; Bray et al. 2019 share a subset of data.

Table 3.8 summarises and compares the C-BARQ factors that had a significant association with training outcomes in assistance dogs in previous studies and this study. Duffy et al. 2012 trained five logistic regression models separately for five different organisations using data collected when the dogs were 12 months old, therefore the number of organisations, for which each factor was significant, is presented. The other studies trained only one model; hence, significant associations between C-BARQ factors and training outcomes are signalled with one, and no significant associations are marked with zero.

Table 3.8: C-BARQ factors significantly related to training outcome in assistance dogs.

| Factors | Duffy et al. 2008 | Duffy et al. 2012 | Bray et al. 2019 | Hunt et al. 2020 | This study |
|---|---|---|---|---|---|
| F_01 Trainability | 0 | 4 | 1 | 0 | 0 |
| F_02 Stranger-directed aggression | 1 | 4 | 0 | NA | 1 |
| F_03 Owner-directed aggression | 0 | 3 | 0 | NA | 0 |
| F_04 Dog rivalry | NA | 3 | 0 | NA | 0 |
| F_05 Stranger-directed fear | 1 | 3 | 1 | NA | 0 |
| F_06 Nonsocial fear | 1 | 4 | 0 | 0 | 1 |
| F_07 Dog-directed aggression | NA | 4 | 1 | NA | 1 |
| F_08 Dog-directed fear | NA | 3 | 0 | NA | 1 |
| F_09 Touch sensitivity | 0 | 3 | 0 | 0 | 0 |
| F_10 Separation related behaviour | 0 | 2 | 1 | 1 | 1 |
| F_11 Excitability | 1 | 3 | 0 | 1 | 1 |
| F_12 Attachment and attention-seeking | 1 | 1 | 0 | 0 | 0 |
| F_13 Chasing | 0 | 2 | NA | 0 | 0 |
| F_14 Energy level | 1 | 5 | 0 | 0 | 0 |

Duffy et al. 2008 reported that C-BARQ scores could discriminate between trainee guide and service dogs that succeeded in training, and those who failed for behavioural reasons. C-BARQ ratings were collected from 5 different assistance dog training organisations when dogs were 6 months and 12 months of age. In particular, the scores assigned to dogs with different training outcomes differed significantly on the following factors F_02 stranger-directed aggression, F_05 stranger-directed fear, F_06 nonsocial fear, F_12 attachment/attention seeking, F_11 excitability, and F_14 energy level.

Subsequently, Duffy et al. 2012 trained 10 logistic regression models as data were grouped by the two age groups and five organisations separately. The models were built on C-BARQ data (14 factors and 22 miscellaneous items) and demographic data (sex and breed) as independent variables, and training outcome as the dependent variable. A total of 27 out of the original 36 C-BARQ items/factors were significantly associated with training outcome at 6 or 12 months of age.

Bray et al. 2019 fitted a generalised linear model with recursive feature elimination using 13 factors and 18 miscellaneous items to predict training outcomes in assistance dogs. The C-BARQ variables with $p < 0.01$ at 12 months were: “I_97 barking”, “F_05 stranger directed fear”, “F_07 dog-directed aggression”, “I_79 coprophagia”, “F_01 trainability”, and “I_85 Pulls excessively hard when on the leash”.

Hunt et al. 2020 compared the predictive ability of eight C-BARQ factors and the correspondent refined puppy walker questionnaire (r-PWQ) trait scores. Only two C-BARQ factors - namely, “F_10 separation-related behaviour” and “F_11 Excitability”- were significantly linked to the outcome as a guide dog compared to five r-PWQ trait scores. No predictive multivariate model was built in this study, but the authors concluded that the r-PWQ could provide a better indication of future success as a guide dog at eight months of age.

Two studies were not included because they did not analyse factors (L. S. Batt et al. 2009; Cleghern et al. 2018). The former modified or combined the C-BARQ items; hence, no comparison can be performed. The latter reported a significant correlation between items I_10, I_11, I_12, and training outcomes in guide dogs. Considering that those items are used for calculating F_02, it could be reasonably estimated that F_02 would have yielded a significant relationship with training outcome if that analysis was performed in that study.

Conclusively, the quantitative comparative analysis of these studies allowed the identification of trends. In particular, some factors were more frequently predictive of training outcomes; namely, “F_02 Stranger-directed aggression”, “F_05 Stranger-directed fear”, “F_06 Nonsocial fear”, “F_07 Dog-directed aggression”, “F_10 Separation related behaviour”, and “F_11 Excitability”. Other factors were only

associated with training outcomes in very few studies; namely, “F_03 Owner-directed aggression”, “F_04 Dog rivalry”, “F_09 Touch sensitivity”, “F_12 Attachment and attention-seeking”, and “F_13 Chasing”. This difference may be due to the low variability in the scores, missing values or factors associated with different organisations such as training methodology and graduation requirements as discussed in Duffy et al. 2012.

#### 3.4.4.2 Prediction Performance

This section compares the prediction performance of the two novel models developed in this study (i.e., the C-BARQ and MCPQ-R models) with previous research. In particular, Duffy et al. 2012; L. S. Batt et al. 2009; Harvey et al. 2017; Cleghern et al. 2018; Bray et al. 2019 also utilised ratings to predict training outcomes in assistance dogs. This comparison revealed that both the C-BARQ and the MCPQ-R achieved a better prediction performance than previous studies. Some studies did not provide important performance metrics, which complicates the comparison. The following paragraphs provide a more in-depth comparative analysis of this and previous studies.

The C-BARQ and MCPQ-R models developed in this research achieved an area under the ROC curves of 0.84 and 0.85 on the cross-validated dataset, which was significantly higher than 0.64 to 0.72 obtained by Duffy et al. 2012 and 0.71 obtained by Bray et al. 2019. Hence, this research advanced the state-of-the-art by providing the most correct machine learning model for the prediction of training outcomes in prospective assistance dogs using the C-BARQ rating.

L. S. Batt et al. 2009 attempted to create the first predictive model for guide dog success using some of the C-BARQ rating items; however, all of them were removed in the recursive elimination. The optimal model contained two variables; namely, the puppy raiser’s estimate of the dog’s chances of success and the number of dogs in the puppy raiser’s household. The former was positively and the latter was negatively related to a successful training outcome. This model achieved a sensitivity of 0.68 and a specificity of 0.63 in predicting a successful outcome.

Cleghern et al. 2018 reports the accuracy, precision, and sensitivity/recall for the model using C-BARQ scores to predict training success. The accuracy achieved by Cleghern et al. 2018’s C-BARQ model was 0.51 while the accuracy achieved by both our C-BARQ and MCPQ-R models was 0.84. Specificity in predicting training withdrawal is mathematically equivalent to sensitivity in predicting training success, as both evaluate the successful dog correctly identified. Hence, the specificity of 0.94 and 0.97 achieved by the C-BARQ and MCPQ-R models, respectively, can be compared to the sensitivity of 0.46 attained by the

C-BARQ model reported by Cleghern et al. 2018. Precision or Positive Predictive Value for successful dogs is equivalent to Negative Predictive Value for withdrawn dogs. Similarly, the NPVs of 0.88 and 0.85 achieved by our C-BARQ and MCPQ-R models, respectively, can be compared to the precision of 0.55 achieved by Cleghern et al. 2018's C-BARQ model. All in all, this indicates that, while Cleghern et al. 2018's model better identified successful dogs, our model provided more accurate predictions of training outcomes.

The prediction performance obtained by our model suggests that the C-BARQ ratings are more accurate for estimating failure than success, which is in line with findings from previous research (Bray et al. 2019). In particular, our C-BARQ and MCPQ-R models achieved a predicted accuracy for the green flags of 91% and 90%, respectively, and for the yellow flags of 100% and 75%, which can be compared to 62% in the top 10% most likely to succeed and 92% in the top 10% most likely to fail in Bray et al. 2019. These results were consistent with the purpose of C-BARQ which was intended to capture potentially problematic behaviours (Hsu et al. 2003; Duffy et al. 2008).

Harvey et al. 2017 collected data on 12 month old dogs utilising the refined Puppy Training Supervisor Questionnaire (PTSQ) to predict the training outcome of guide dogs. A flagging system based on a composite z-value was proposed for the prediction of training outcomes. The predictive value for the green flag was 85%, while our C-BARQ and MCPQ-R models' green flags achieved 100% and 75%, respectively. The sensitivity for the green flag was 8.4% which was much smaller than the precision achieved by the green flag of our C-BARQ and MCPQ-R models of 94% and 90%, respectively. The predictive value for the red flag was 83%, while our C-BARQ and MCPQ-R models' yellow flags achieved 91% and 90%, respectively. The sensitivity for the red flag was 8.4% which was much smaller than the sensitivity achieved by the yellow flag of our C-BARQ and MCPQ-R models of 36% and 33%, respectively.

# Chapter 4

# Assistance Dog Test Battery (ADBT)

## 4.1 Introduction

The work presented in this chapter aims to achieve the primary objective of this thesis which is to evaluate and compare the validity of traditional and novel behaviour assessment methods to predict working outcomes for trainee assistance dogs. The most commonly used methods are ratings, which are utilised in Chapter 3, and test batteries, which are employed in this chapter. Test batteries present significant advantages over ratings as they do not rely on the rater's familiarity with the subject dog, their subjective judgement, and/or their ability to recall recent behaviour.

This study builds upon the current state-of-the-art by combining subtests used in previous research into a novel test battery, to assess a trainee's dog suitability for assistance work. The purpose of this study was to investigate the use of behavioural analysis from a novel test battery, for enabling the early prediction of future working outcomes in an assistance dog programme. The main goal was to build a statistical model, capable of estimating a dog's probability to be withdrawn either during training or during the first year of the working programme. This research benefits assistance dog providers, as the early prediction of working outcomes allows them to withdraw the least suitable trainee dogs earlier, reducing costs, and increasing efficiency.

Test batteries comprise a subtest protocol and an ethogram marking system. Taylor et al. 2006 identified and discussed five key measures of quality of a test battery; namely, purpose, standardisation, reliability, validity, and practicality. The purpose of the test, standardisation and practicality were considered for the design of the test battery. Results from previous research in this area were thoroughly evaluated to select

relevant subtests and marking methods, which were adapted and combined into a novel test battery.

Taylor et al. 2006 and Brady et al. 2018 highlighted the importance of evaluating canine behaviour tests based on validity and reliability tests. This present initial study focuses on the former; hence, the validity of the novel test battery was computed using the subtest protocol and ethogram marking system with a selection of prospective assistance dogs. The latter will be analysed in a future study when a sufficient number of repeated ethograms has been collected.

Predictive validity can be assessed considering two measures of success in working dogs; namely, graduating success and working success (L. Batt et al. 2010). Graduating success (also referred to as training success or training outcome in the literature) is the number of dogs graduating and is the most commonly used measure, as it requires less time to define. On the other hand, working success is the number of dogs still working with clients one year after graduation. It incorporates more variables than graduating success as it is affected by the strength of owner attachment and stability of temperament. Working success is considered a more definite measure of true success than graduating success, as the latter is affected by the subjective assessment by different trainers and external factors related to supply and demand for assistance dogs.

Conclusively, the purpose of this study was to design and evaluate the validity and predictability of a novel test battery for assessing trainee assistance dogs at different stages of their training. The proposed test was entitled Assistance Dog Test Battery (ADTB). The main goal was to develop the most reliable machine learning model for estimating working outcomes in trainee assistance dogs. It was also of interest to determine the optimal age for evaluation, in order to improve predictions of working outcomes. Accordingly, the following research questions were investigated:

**Q.4** Is the subtest protocol and ethogram marking system developed for evaluating the novel ADTB valid for assessing the working outcome?

**Q.5** Can machine learning models predict working success using variables from the ethogram items and composite scores developed for evaluating the novel ADTB?

**Q.6** When is the optimal time to perform the novel ADTB on trainee dogs in order to maximise working success prediction performance?

## 4.2 Methodology

In order to achieve the above aims and answer the associated research questions, Section 4.2.3 introduces the subtest protocol and the ethogram scoring system developed for the novel ADTB. Section 4.2.4 describes the test arena and room layout used to carry out the test battery. Subsequently, Section 4.2.5 presents the methods used to evaluate the validity and predictability at two different stages of training considering the training outcome.

### 4.2.1 Ethical Approval

The novel ADTB subtest protocol and ethogram scoring procedures were reviewed and approved by the Animal Experimentation Ethics Committee (AEEC) and the Social Research Ethics Committee (SREC) at University College Cork (UCC), as described in this section under request number 2019-007 and 2019-016, respectively. Dog trainers participating in this research were briefed on the purpose of the research and the scope of their contribution. Following their consent, they assisted the test leader responsible for the data collection by playing the role of the handler during the test battery.

### 4.2.2 Subjects

The target number of subjects for this research was 100 dogs (Laborde et al. 2017). Data collection was affected by the global COVID-19 pandemic causing a significant delay in this part of the project.

The trainee dogs participating in this research were Labrador Retriever (LR) and Labrador Retriever crossed with Golden Retriever (LRxGR) which are the most common and successful breeds of assistance dog (Ennik et al. 2006; Duffy et al. 2012). Only two dogs from other breeds were included; in particular, they were one Golden Retriever (GR) and one Golden Retriever crossed with a Poodle.

Dogs included in this research were either withdrawn from the training or working programmes for behavioural reasons, or worked with a blind or disabled person for at least one year. Henceforth, the former group are referred to as “fail” and the latter as “success”. Dogs who were withdrawn from the programme due to health issues or selected as breeding stock were removed from the sample.

### 4.2.3 Test Battery

The novel Assistance Dog Test Battery (ADTB), including the subtests and ethogram, was designed in line with previously published research

results and adapted considering constraints, research goals and feedback from researchers and practitioners supporting this research. Specifically, this novel test battery was developed based on the relationship between behaviours presented by trainee assistance dogs during test batteries, and their future training outcomes. Two review papers presented relevant research that supported the development of this novel test battery (Taylor et al. 2006; Marcato et al. 2022) - see also Chapter 2.

Taylor et al. 2006 reviewed five studies using test batteries to evaluate the temperament of companion dogs and found common features between those and tests developed to assess the suitability of assistance dogs. This similarity was attributed to the fact that both tests share the same overarching theme of identifying a friendly and willing companion. They identified and summarised 20 subtests typically utilised. Chapter 2 examined ten peer-reviewed studies on test batteries used to assess assistance dogs. Eight studies analysed the relationship between the items recorded and training outcomes. The items showing a statistically significant association with training outcomes were highlighted.

Taylor et al. 2006 advised researchers to carefully cogitate on the need for subtests that provoke fear, anxiety, or aggressive behaviour, and encouraged them to investigate the trade-off between maximising predictability and minimising stress experienced by the animal. Brady et al. 2018 concluded that there was a clear relationship between positive affect behaviour response and performance in guide dogs, while less consensus was found on the validity of negative affect behaviour response for predicting training outcomes in working dogs. Hence, subtests for assessing positive affect were included as they offer a great opportunity to evaluate trainee assistance dogs while eliminating animal welfare concerns.

In this thesis, the novel test battery protocol proposed was entitled Assistance Dog Test Battery (ADTB). It was created with the purpose of differentiating dogs with different working outcomes. Table 4.1 provides details of the subtests' order, name, description, and references to the studies they were adapted from. This novel test battery comprised 14 subtests, which were divided into 5 sections in the following order: activity, responsiveness, sensitivity, focus, and emotional regulation.

Table 4.1: Detailed description of the protocol for data collection using the ADTB.

| **Subtest (Duration)** | **Description (References)** |
|---|---|
| | **Part I. Activity** |

Continued on next page.

Table 4.1 - continued from previous page.

| **Subtest**<br>**Duration** | **Description**<br>**References** |
|---|---|
| 1. Familiarisation<br>(1min) | Handler and dog entered the test arena, handler let the dog off-leash, the dog was allowed to explore the arena for 1 minute, handler called the dog back if they left their side (Wilsson et al. 1998) |
| **Part II. Responsiveness** | |
| 2. Walking<br>(1min) | Handler walked dog on a leash (Wilsson et al. 2012) |
| 3. Standing<br>(1min) | Handler commanded the dog to “stand” and “wait”; if dog left the position, the handler would command them again until 1 minute had elapsed (Harvey et al. 2016b) |
| 4. Sitting<br>(1min) | Handler commanded the dog to “sit” and “wait”; if dog left the position, the handler would command them again until 1 minute had elapsed (Harvey et al. 2016b) |
| 5. Lying<br>(1min) | Handler commanded dogs to lie down and stay using “down” and “wait”; if the dog left the position, the handler would command them again until 1 minute had elapsed (Harvey et al. 2016b) |
| **Part III. Sensitivity** | |
| 6. Body check<br>(1min) | Handler and test leader encouraged the dog to jump up on the table; if they would not go on the table, the physical examination was carried out on the floor; the test leader gently restrained the dog by wrapping their arms around their chest and abdominal circumferences; handler carried out physical examination checking their eyes, mouth, ears, paws, tail (on leash) (Asher et al. 2013; Harvey et al. 2016b; Bray et al. 2019) |
| 7. Tea Towel<br>(1min) | Handler placed a quarter folded tea towel over the dog’s back (off leash) while they were standing; handler waited for 5 seconds or until the dog removed the towel, repeating this two times (Harvey et al. 2016b) |
| **Part IV. Focus** | |
| 8. Distractions<br>(2min) | Handler distracted dog facing the back of the room while test leader placed the objects in a circuit around the test arena; Handler and dog (on leash) walked past the different stimuli, repeating this two times in the following order:<br>(1) Teddy - teddy bear<br>(2) Human - test leader greeted dog (Svartberg et al. 2002; Bray et al. 2020)<br>(3) Car - remote-controlled car moved along the path<br>(4) Food - the bowl contained dog food (Harvey et al. 2016b; Bray et al. 2017b) |

Continued on next page.

Table 4.1 - continued from previous page.

| **Subtest Duration** | **Description References** |
|---|---|
| 9. Kong (2min) | (1) Presentation - The handler held the Kong (Kong, Golden, Colorado, United States of America) at chest height and commanded the dog (off leash) to sit and wait in front of them for 40 seconds (2) Interaction - The handler gives the Kong to the dog (off leash), and the dog is allowed to interact with Kong for 20 seconds (3) Return - The handler takes the Kong away from the dog (off leash) and returns it to storage (L. Batt et al. 2008a; L. M. Tomkins et al. 2012; MacLean et al. 2018) |
| 10. Dog distraction (30sec) | Handler commands the dog to sit and wait (off leash) by its side, then the test leader gets the distraction dog from a crate and walks them by the fence; the dog is allowed to behave naturally until the distraction dog enters their crate; then handler calls the dog back if they had left their side (L. Batt et al. 2008a; L. M. Tomkins et al. 2011; Bray et al. 2019) |
| | **Part V. Emotional Regulation** |
| 11. Crate (1min) | Handler encouraged the dog to enter the crate; in the case where the dog would not go in, this subtest was skipped; otherwise, the handler closed the crate once the dog was in and sat on the chair ignoring dog for 1 minute; handler opened the crate allowing the dog to come out (Svartberg 2002; L. Batt et al. 2008a; L. M. Tomkins et al. 2011) |
| 12. Petting (1min) | Handler crouches down, freely petted and played with a dog (off leash) for about 20 seconds; handler would sometimes hold the dog by the collar to ensure physical proximity and encouraged them to engage with the play (Asher et al. 2013) |
| 13. Isolation (1min) | Handler left dog (off leash) alone in the test arena; both handler and test leader left room for 1 minute; handler and test leader returned to room; handler greeted, petted and attached the leash on dog's collar (Wilsson et al. 1998; Bray et al. 2017b) |
| 14. Noise (1min) | Handler walked the dog (on leash) around the test arena; test leader stood by the fence holding a metal bucket on their first round; test leader dropped the metal bucket from knee height in their second round when the dog was approximately 2 meters away from and looking at them (Wilsson et al. 1997b; L. Batt et al. 2008a; L. M. Tomkins et al. 2011; Wilsson et al. 2012; Asher et al. 2013; Bray et al. 2019) |
| **Estimated Total Duration: 16min** | |

Continued on next page.

Subtests were combined into sections with an overarching purpose as shown in Table 4.1 and were presented to the subject dog in a controlled environment following a predetermined order. Part I captured the subject dog's response to a novel environment. Part II was intended to gather inertial data for the posture recognition system and behavioural data for examining responsiveness to basic commands. Part III evaluated body sensitivity and Part IV evaluated distractibility and ability to focus. Part V assessed emotional regulation characteristics such as fearfulness, aggressiveness, adaptability, and separation-related behaviours.

Key elements of 12 subtests presented in Taylor et al. 2006 were incorporated into 13 subtests in the ADTB, as the description of the "novel room test" applies to both the "familiarisation" and "isolation" subtests. The additional eight subtests presented by Taylor et al. 2006 were not considered for practicality purposes ("approach dog in the kennel", "doll test"), welfare concerns ("threatening approach", "umbrella test", "skin sensitivity"), and lack of evidence for assistance dog assessment ("stare at the dog, collar", "food guarding"). The only subtest derived from a different study was the "tea towel" test, which came from (Harvey et al. 2016b). Elements from all the ten test batteries summarised in Chapter 2 appeared in the subtests of the ADTB.

In order to eliminate welfare concerns, stimuli were carefully selected to prevent potential strong fear responses. Moreover, the duration of the test was kept to around 16 minutes. The subtests replicate a collection of typical real-life situations which are recurrent in the life of an assistance dog. They included both positive and negative stimuli given that preliminary evidence of a relationship between such stimuli and training outcomes existed.

An ethogram scoring system was developed to rate the dog's behavioural reaction considering previous research and dog trainers' feedback. Table C.1 in Appendix C.1 describes the final ethogram including information on the behaviours assessed and the scoring system utilised. References are provided for the measures that were created based on other studies.

The next paragraphs further discuss the novel test battery ethogram and the studies from which the subtests were adapted. They describe the main similarities and differences between them while highlighting the novelty in the ADTB.

Part I allowed the subject dog to familiarise itself with the new environment on their own terms as they were off-leash. This is similar to the "arena test" in Wilsson et al. 1998, except that no objects were on

the arena other than the permanent furniture of the room. Subsequently, in subtest 2 “walking”, the dog explored the environment on a lead, guided by the handler, as they walked around the test arena. This is similar to the following subtests in three reported studies: (1) the “leash” subtest in Wilsson et al. 2012 performed by prospective military dogs, except that changes in direction were not as frequent; (2) the “walking on leash” subtest in Lucidi et al. 2005, performed by shelter dogs to evaluate their suitability for service work; and (3) the “on leash behaviour” subtest in Weiss et al. 1997.

Part II comprised subtests 2 to 5 which were designed to gather inertial data for the posture recognition system presented in Chapter 5. Additionally, the behavioural response displayed by the dog on these subtests was also analysed to assess obedience and trainability. Subject dogs were commanded to stand in subtest 3 for inertial data collection purposes, as this is not a command that assistance dogs are normally trained on. Subtests 4 and 5 are similar to the “obedience” subtest in Harvey et al. 2016b, except that the handler was the dog’s trainer instead of their puppy raiser or a stranger, and “obedience training” in Weiss et al. 1997; Weiss 2002.

Part III was designed to assess the dog’s level of body sensitivity. Subtests 6 and 7 were inspired by the work published by Harvey et al. 2016b. In order to further investigate the relationship between body sensitivity and training outcome in 5 and 8 months old prospective guide dogs, Harvey et al. 2016b carried out three body sensitivity subtests; namely, “body check”, “head ring”, and “tea towel”. “Head ring” was not associated with training outcome; hence, it was not included in this research. Ethogram scores associated with the “body check” subtest were not significant in Harvey et al. 2016b, however, Bray et al. 2019 demonstrated promising results using different ethogram measures. Therefore, it was included as subtest 6 “body check” in this research. Finally, as ethogram measures associated with the tea towel achieved promising results, it was replicated in the exactly same fashion in subtest 7 “tea towel”.

Part IV evaluates the dog’s ability to focus, which can be considered the opposite of distractibility. Arata et al. 2010 demonstrated an association between distraction and guide dog qualification, and suggested that behavioural tests incorporate subtests to objectively evaluate distraction earlier than 3 months into the training programme. Harvey et al. 2016b proposed placing different distracting stimuli along a route through which the subject dog was led on a leash by the handler, in the following order: “food” (paper plates with sausages), “robin” (remote-controlled toy), “pigeons” (plastic pigeon toy) and “human distraction” (researcher stood still and looked at the dog). If the dog stopped walking or tried engaging with any of the stimuli, the handlers would walk them back to the start of the circuit until they moved on.

Accordingly, subtest 8 “distractions” was designed to assess the dog’s ability to focus on walking in the face of different distracting stimuli (food, people, inanimate and animate objects). This process was repeated twice to enable the evaluation of the change in reaction. It was adapted from Harvey et al. 2016b in order to simplify the materials used and shorten the duration of the test for practical use. The former was achieved by using a bowl of dog food instead of sausages, a standard remote-controlled toy car instead of a robin, and a teddy bear instead of pigeons. The latter was addressed by instructing the handler to walk the subject dog from the start point around this circuit twice, continuing on the route instead of going back to the start. The handler used commands such as “come on” or “leave it” followed by the dog’s name whenever they got distracted but continued the route instead of returning to the start point.

Subtest 9 “Kong” was adapted from the subtests “Kong test” in L. Batt et al. 2008a and L. M. Tomkins et al. 2012, “first-stepping” in L. M. Tomkins et al. 2012 and “laterality”, as they found a significant relationship between guide dog training success and measures of laterality. The Kong toy used in this subtest was filled with a mixture of highly palatable foods prior to freezing it. The calculation of the laterality index as outlined L. Tomkins et al. 2010 requires 50 valid data points between left or right paw uses. However, for practical use, it can be time-consuming and difficult to determine when 50 data points are completed. Moreover, another significant challenge is to capture valid data points in the video footage, due to the dog’s body obscuring the camera view or the dog not being in the shoot, as reported in L. Batt et al. 2008a. Hence, this research investigated a more practical version of this subtest for the first time.

Subtest 10 “dog distraction” is similar to the “dog distraction test” in L. Batt et al. 2008a and L. M. Tomkins et al. 2011, as well as the “unfamiliar dog test” in Bray et al. 2019. The occurrence of jumping and pulling in L. Batt et al. 2008a, the occurrence of painting and the number of licks in L. M. Tomkins et al. 2011 during the dog distraction test were significantly associated with a lower probability of graduation in trainee guide dogs. Bray et al. 2019 discovered a significant relationship between scores in the recall, initial reaction, tail position and barks or growls in the unfamiliar dog subtests, and training outcomes in assistance dogs, at different points of the training programme.

Part V assesses emotional regulation by analysing behavioural reactions in response to stimuli and situations which can potentially induce change in emotional state or stress (Katayama et al. 2016; Asher et al. 2013; Maros et al. 2008). Subtest 11 “crate” was modified from the “passive situation” in Svartberg 2002 and “passive test” in L. Batt et al. 2008a and L. M. Tomkins et al. 2011. The test also involved the handler sitting on a chair and ignoring the dog; however, two main adjustments

were made. Firstly, the dog was placed in a crate about 3 metres from the handler instead of being leashed. Secondly, the test duration was shortened to 1 minute; unlike in Svartberg 2002 where the test duration was limited to 3 minutes or L. Batt et al. 2008a, and L. M. Tomkins et al. 2011 where the test ended either when the dog rested or after 6 minutes. These significant differences in the subtest protocol and ethogram measures originated a novel subtest.

Subtest 12 "petting" was adapted from the "stroking" stimulus proposed in Asher et al. 2013. This subtest aimed at observing the engagement between dog and handler in a somewhat spontaneous way. Asher et al. 2013 found that trainee dogs who scored lower in response to the assessor were significantly less likely to graduate as guide dogs. In this study, this subtest incorporated elements of both petting and physical play, as handlers were allowed to interact with the dogs in their own way providing a more natural demonstration of the handler-dog bond.

Subtest 13 "isolation" was similar to the "isolation" test in Wilsson et al. 1998 and Bray et al. 2017a. It intended to capture the emotional distress caused by being isolated. This study differed from those previously published in two aspects: order and duration. In past studies, the isolation test was the first one in the test battery; therefore, the dog was not familiar with the room in that instance. In this study; however, the isolation subtest was the second last in the test battery; hence, the dog was familiarised with the test arena and the main stimuli were being left alone. Lastly, the duration of the subtest was reduced from 2 minutes to 1 minute.

Subtest 14 "noise" was based on a combination of the "loud noise" test in Wilsson et al. 1997b, "metallic noise" in Svartberg et al. 2002, "noise test" in L. Batt et al. 2008a; L. M. Tomkins et al. 2011 and Asher et al. 2013, "acoustic startle" in Wilsson et al. 2012, "sudden noise" in Bray et al. 2019. This test can be considered one of the most popular in behavioural research with assistance dogs. The method used for generating a loud noise varied, for instance, Svartberg et al. 2002 and Bray et al. 2019 dragged a heavy metal chain while the dog was in close proximity, L. Batt et al. 2008a; L. M. Tomkins et al. 2011 dropped a plate in an adjacent room while the dog was in the test arena and Asher et al. 2013 played a CD or aircraft noise through speakers. In particular, the design of this subtest was inspired by Wilsson et al. 1997b and Wilsson et al. 2012, who dropped metal buckets while the dog walked on a leash. Unlike Wilsson et al. 2012, the handler did not drop the leash and encourage the dog to investigate the bucket; instead, they encouraged the dog to keep walking commanding "leave it" or "come on" and calling the dog's name.

A pilot study was conducted to define an optimal order for the subtests which is presented in Table 4.1, considering test requirements and the

typical emotional response elicited. In terms of requirements, measures were taken to minimise the number of people (test leader and handler) and dogs (subject dog and distraction dog) involved. The dog's emotional response was important because it affected their engagement in the next subtest. For instance, at the end of the distractions subtest, a bowl of food is shown which generally excites the subject dog. This event can keep distracting the dogs, especially in the following subtest.

Some measures were also taken to ensure a high level of standardisation and reproducibility of the procedures with the purpose of preventing or minimising the effects of potential miscellaneous factors. For instance, the test leader briefed the handler on the test battery protocol using an information poster, while indicating the location of relevant subtests. The poster describing the order and key information about the subtests was hung on the wall, remaining visible to the handler at all times during the test battery. The test leader kept a sheet describing the order and key information on each subtest in hand at all times as well as an iPad (Apple Inc., Cupertino, California, United States), to keep track of the duration of the subtests. Moreover, the exact location of some relevant subtests was physically marked with tape on the floor.

In order to investigate the optimal age for training, subject dogs participating in this research underwent the same test battery twice, considering their date of arrival at the training centre for formal training: the first instance took place at 3 weeks of formal training, and the second instance occurred at 10 weeks. Henceforth, they were referred to as Data Collection 1 (DC1) and Data Collection 2 (DC2), respectively.

The test battery data collections followed the subtests protocol and were video recorded to permit a subsequent detailed behavioural analysis. Two experienced assistance dog trainers at the IGDB carried out the behaviour analysis by watching the video recording and rating the dog's reaction using the ethogram.

In addition to the behavioural assessment, inertial, and physiological data were also monitored. The inertial data was collected by three Actigraph GT9X Links. The physiological data was acquired by the Polar H10 and Bioharness 3 which monitored cardio-respiratory parameters ECG, IBI, and BR. Analysis of these data is outside the scope of this thesis; however, the datasets will be made publicly available providing a valuable contribution to the advancement of the field.

Conclusively, the ADTB builds upon the current state-of-the-art by combining subtests from a number of published studies into one systematised and standardised test battery for the first time. The design process considered not only evidence from previous research in this area, but also concerns raised by the IGDB and guiding principles presented in Taylor et al. 2006, including purpose and duration of the

test, dog welfare concerns, and feasibility for practical use.

### 4.2.4 Test Arena

The test battery was conducted at the IGDB training centre located in Cork, Ireland between 11.30 am and 4 pm. It was carried out by a test leader and a handler, who were normally a researcher and the dog's trainer, respectively. The test leader was responsible for starting the camera and directing the handler where appropriate. Their responsibilities during the test battery are outlined in Table 4.1. On some occasions, the test leader was accompanied by another researcher who acted as a test assistant, performing some of the test leader's responsibilities.

Figure 4.1 displays a schematic representation of the experimental set-up for the test arena, with the subtest numbers displayed accordingly. The subject dog was let off leash during subtests 1 and 13; hence, these are not indicated. The test arena was set up in a familiar room.

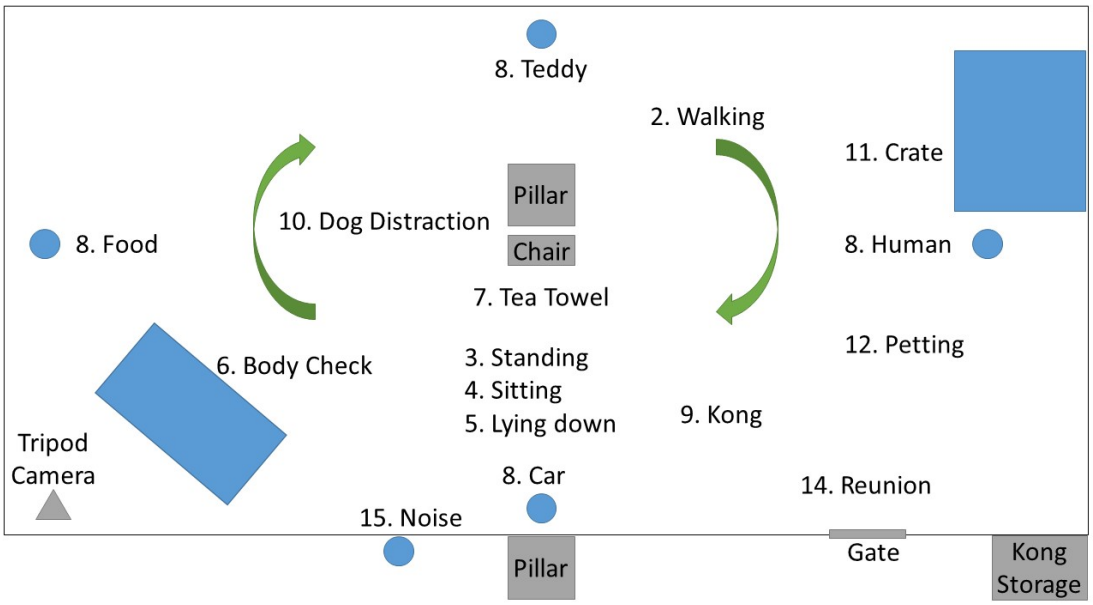



Figure 4.1: Schematic representation of the test arena including key materials used and information about the test battery.

Figure 4.2 presents the materials used in the test battery. In order to minimise requirements and promote reproducibility, the material list was simplified to include only widely available, easily accessible or commonly used objects in dog training facilities.

An additional dog was necessary for the "dog distraction" subtest; therefore, a dog available at the kennels was used. The main criteria used for selecting the dog related to their ability to settle quietly in a

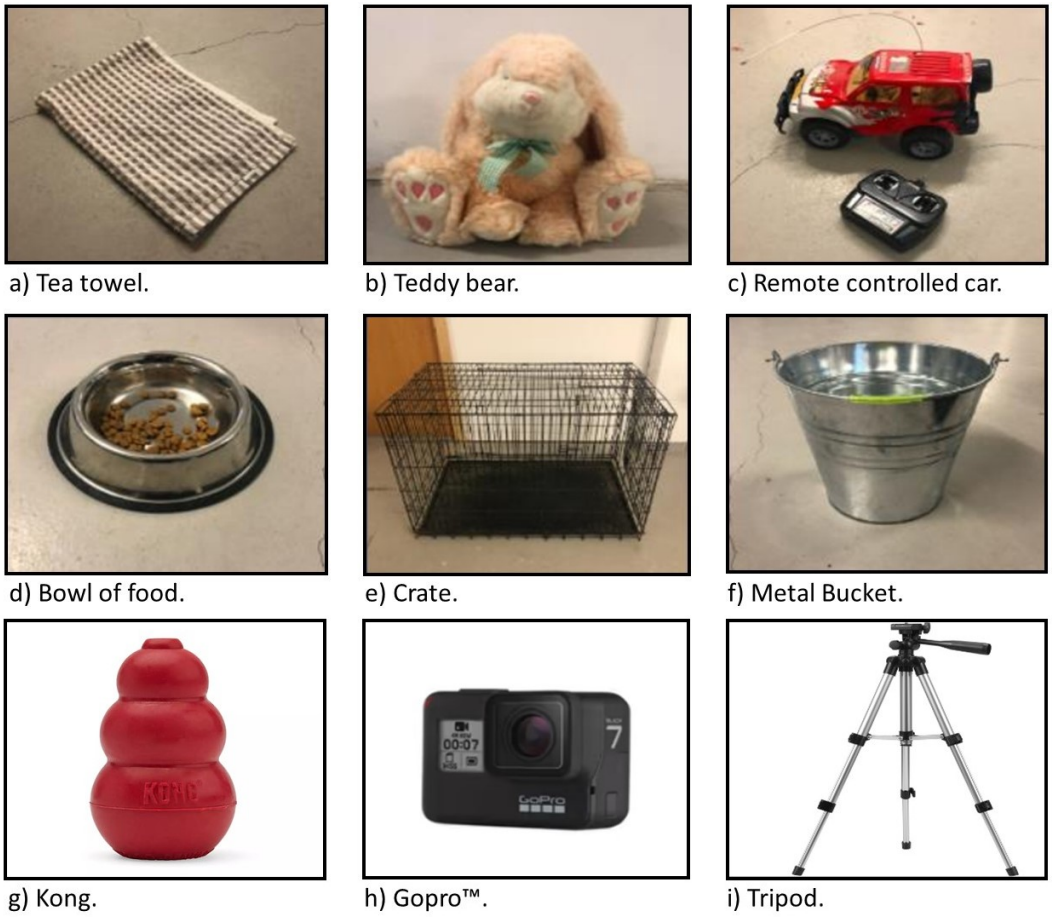



Figure 4.2: Materials used in the test battery.

crate for a couple of hours. The distraction dog was kept in a crate hidden from the subject dog in the experimental room outside the test arena. The distraction dog was given a treat to chew on and was checked regularly while they rested waiting in the crate.

### 4.2.5 Data Analysis

The data were analysed using R version 4.2.1 (Team 2022). Generalised Linear Models with a logit link function were used to implement the logistic regression models, for both feature selection and prediction of training outcome (Hosmer et al. 2013). They are referred to as logistic regression models for brevity henceforth. Unless otherwise stated, significance was set at $P < 0.05$.

The working outcome was a categorical variable where success was assigned to dogs who completed the training as either guide, assistance or community dogs and worked with a blind or disabled person for at least one year; and fail was assigned to dogs who were withdrawn from either the training or working programmes for behavioural reasons.

The original variables included in the dataset belonged to two sources: demographic characteristics (N = 2 variables; sex and breed) and ethogram items (N = 70 variables; as detailed in Table C.1).

Composite scores were calculated based on ethogram items in order to combine information from multiple items into one variable. Appendix C.2 contains a detailed description of the calculation of the composite scores (N = 41). The resulting dataset considered for analysis contained a total of 111 features; specifically, they were 17 categorical and 53 ordinal ethogram items, and 41 derived numeric scores. For the purpose of the data analysis, the variables were classified into two types; namely, ordinal and categorical variables. Accordingly, they were converted to numeric and factor data types in R, respectively.

In order to investigate the most optimal stage of training to test trainee dogs, separate analyses were carried out with the data collected at 3 weeks (DC1) and 10 weeks (DC2) of training independently. The exact same methodology was followed for the data analysis, the only difference was the timing of testing. In particular, the evaluation of validity and predictability was conducted using data collection 1 (3 weeks of training) and data collection 2 (10 weeks of training) considering the methodology presented in Sections 4.2.5.1 and 4.2.5.2.

#### 4.2.5.1 Validity

The validity of the ethogram was evaluated by analysing the level of association between the variables recorded and the training outcome. Univariate logistic regression models were built to test the relationship with training outcomes, considering each of the ordinal and categorical ethogram variables individually, as well as the composite scores calculated from those original items. The following three levels of significance were considered: $p < 0.05$, $p < 0.10$ and $p < 0.15$.

#### 4.2.5.2 Predictability

The resulting dataset consisted of 111 ADTB ethogram variables, including 70 original items (53 ordinal and 17 categorical) and 41 composite scores, which were calculated based on the original items. The feature selection strategy utilised in this study was adapted from the purposeful selection of features methodology proposed by Hosmer et al. 2013, and is similar to research in this area by Asher et al. 2013 and Harvey et al. 2016b. The following three criteria were used to remove features less predictive of working outcome:

1. **Step 1 - univariate models**: fit a univariate logistic regression model for each feature (independent variable), using one feature as a predictor, and working outcome as the desired label (dependent

variable). The threshold for retention was set at $p < 0.15$; therefore, variables with a higher p-value were removed.

2. **Step 2 - independent variables**: variables associated with each other in two different scenarios were removed as described below: (1) where both the composite scores' product and mean, derived from the same original ethogram items met the threshold for retention ($p < 0.15$), the score with the lowest p-value was retained; (2) where both the derived composite score and the original ethogram items used to calculate it met the threshold for retention ($p < 0.15$), and the variable with the lowest p-value was retained.

   This approach was adopted to select covariates more strongly related to training outcomes and avoid multicollinearity in the multivariate model.

3. **Step 3 - correlation**: analysis of the pair-wise correlation between the predictor variables remaining for consideration after the previous step. Whenever the correlation between a pair of numerical covariates was above 0.40, one of them was removed to prevent multicollinearity issues in the multivariate model. The variable presenting the highest mean absolute correlation with all other covariates was eliminated. Whenever there was a significant association between two categorical covariates as tested by the Chi-square test, the variables with the weakest relationship with the target variable (highest p-value) were removed.

The reduced feature dataset was pre-processed in two steps: (1) imputation of missing values with the mean, and (2) normalisation of values using the z-score. Subsequently, multivariate logistic regression was used to model the relationship between the independent variables in the reduced dataset composed of $M$ features, and the dependent variable training outcome using $N$ examples.

In order to maximise the number of training examples, the sampling strategy was Leave One Out Cross Validation (LOOCV). The LOOCV algorithm fitted the model to the entire dataset excluding the example used for validation ($N - 1$) until all $N$ examples were used for validation. The predictions made using the $N$ validation examples were then used to calculate cross-validated performance metrics.

The model predicted the probability of failure instead of success; hence, the case or positive class is "failure", and the control or negative class is "success". Model performance was optimised to maximise the f1-score which is defined as the harmonic average of the precision and recall. It varies between the worst score of 0 and the best score of 1. The f1-score was selected as the optimisation metric because it measures the model's ability to correctly predict both the positive and negative classes.

A backward feature selection algorithm was implemented using

Recursive Feature Elimination (RFE) to maximise the f1-score. This RFE algorithm started by fitting a model using all features in the reduced dataset ($M$) to create a ranking for the features; then it removed the lowest ranked feature at each iteration until only one feature was left (i.e., $M, M-1, ..., 2, 1$). The model which achieved the highest f1-score was selected as the optimal model.

Model fitness was analysed using the likelihood ratio test for nested models, which follows a chi-squared distribution. The likelihood ratio test statistic was calculated based on the difference between the deviance of the null model and proposed model in Equation 4.1, and the degrees of freedom equal to the reduction in the degrees of freedom as shown in Equation 4.2. The proposed model was considered a good fit, if the p-value resulting from the chi-squared test performed with the critical value $\chi^2 = D_{Difference}, df = df_{Difference}$ was smaller than the significance level of 0.05 (Hosmer et al. 2013).

$$D_{Difference} = D_{Null} - D_{Proposed} \tag{4.1}$$

$$df_{Difference} = df_{Null} - df_{Proposed} \tag{4.2}$$

Precision, sensitivity, and specificity are one-dimensional or partial performance metrics since they only take into account the results from either the positive or the negative class, but not both (Luque et al. 2019). It is important to estimate the performance of the optimal model using a metric that takes into account this trade-off. In particular, the PR curve considers the trade-off between precision and recall/sensitivity, and the ROC curve is a measure of sensitivity/recall and specificity (Jiang et al. 2008).

The areas under the PR and ROC curves provide a single measure, reflecting the balance between metrics. Their values range between 0.5 and 1, with a value of 0.5 indicating a non-informative model, and a value of 1 indicating a perfectly predictive model. The area under the ROC analysis does not have any bias toward models that perform well on the minority class, at the expense of the majority class, which makes it quite attractive when dealing with imbalanced data (Luque et al. 2019). It is reported as the most commonly used metric for imbalanced datasets (Haixiang et al. 2017).

Conclusively, the f1-score was used to optimise the model, and the area under ROC curve is used to validate and compare models. Precision, sensitivity, specificity, recall, and area under the PR curve are also reported. All metrics were calculated using the predictions on the validation set generated with LOOCV.

The estimates from the optimal multivariate logistic regression model were analysed through the calculation of the Odds Ratio (OR) and 95% Confidence Interval (CI). The interpretation of the optimal model is limited because all covariates are composite scores which were normalised before learning. The impact that the features in the optimal model have on the final predictions was analysed. The feature importance of the logistic regression model was estimated using the absolute value of the t-test.

## 4.3 Results

### 4.3.1 Data Collection 1

The sample comprised a total of 98 dogs, including 49 females (50%) and 49 males (50%); there were 43 Labrador Retriever, 53 Labrador Retriever crossed with Golden Retriever and 1 Golden Retriever and 1 Golden Retriever crossed with Poodle. Their age at arrival ranged from 12 to 22 months (M = 15.3, SD = 2.04). Their age at Data Collection 1 varied from 13 to 23 months (M = 16.15, SD = 2.01). The failure rate in this sample was 21%, as 21 dogs failed and 77 dogs succeeded in their training.

Figure 4.3 shows the duration of training for the dogs withdrawn from training for behavioural reasons. Dogs were withdrawn at 15 weeks of training on average. A total of eight out of the 21 dogs were withdrawn before 10 weeks, while only one was withdrawn before 3 weeks. Only one dog was withdrawn during the first year of the working programme.

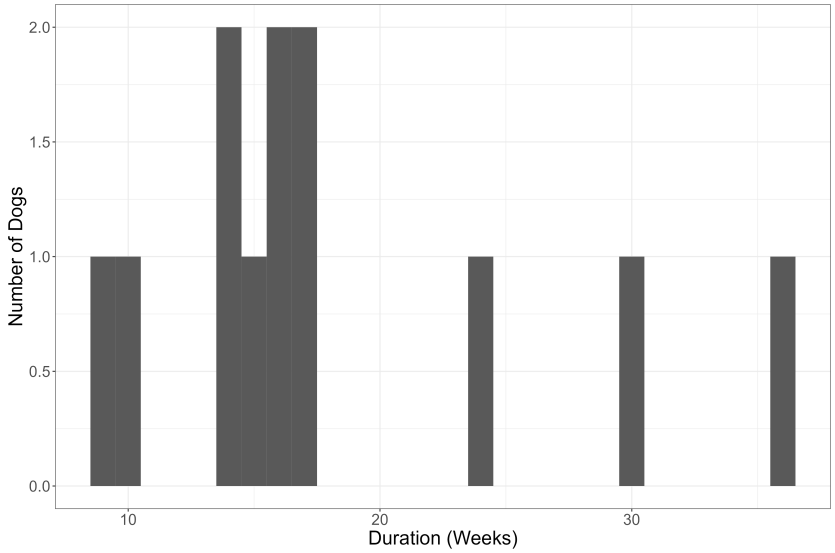



Figure 4.3: Duration of training in weeks for dogs that were withdrawn from training.

Table C.3 shows descriptive statistics of the numerical variables in this sample including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing data of all original items and composite scores. Table C.4 describes the number of instances per category and missing data in this sample. The marking scale used to score the items is described in Table C.1.

#### 4.3.1.1 Validity

Univariate logistic regression was used to investigate whether there was a statistically significant difference between the working outcome of the dogs and ADTB ethogram items, composite scores, and their demographic information. The results from the logistic regression univariate models are shown in Table C.7, highlighting three levels of significance. A total of two variables had $p < 0.05$, nine were in the range $0.05 < p < 0.10$ and five were within $0.10 < p < 0.15$. The categorical variables "Breed" and "Kong Interaction Lateralisation" had more than two categories, thus their results are shown in Tables C.8 and C.9.

Only two variables from the ADTB showed a significant association with working outcomes at traditional levels ($p < 0.05$). They were the ethogram item "Petting Confidence During" ($p = 0.03$) and its derived composite score "S Petting Stimulus During_mean" ($p = 0.02$). Normalisation was used before fitting the model to the data; hence, the scales changed in each fold depending on the average and standard deviation of the sample in each training set. Therefore, quantitative analysis is limited, as the level of change in behaviours cannot be fully determined.

The OR for "Petting Confidence During" was 0.51, suggesting that a unit increase of the level of confidence decreases the odds of failure by (0.51 - 1 = -0.49) 49%. There is a 95% probability that the level of decrease in the odds of failure lies between (0.91 - 1 = -0.09) 9% and (0.28 - 1 = -0.72) 72%. Similarly, a unit increase of "S Petting Stimulus During_mean" decreases the odds of failure by (0.27 - 1 = -0.73) 73%. There is a 95% probability that the level of decrease in the odds of failure lies between (0.75 - 1 = -0.25) 25% and (0.08 - 1 = -0.92) 92%. "S Petting Stimulus During_mean" was calculated by taking the mean of "Petting Handler Stimulus", and "Petting Confidence During" as shown in Table C.2 to reflect the engagement between handler and dog.

#### 4.3.1.2 Predictability

A three-step feature selection strategy was carried out. The initial dataset contained two demographic variables, 70 original items and 41 composite scores. Table 4.2 indicates which of the variables selected in step 1 satisfied steps 2 and 3, and were selected by the RFE algorithm.

They were arranged according to the order of the subtests as shown in Table 4.1 and the lines delineate groups of variables that are related to each other, as indicated in Table C.2.

Table 4.2: Variables that satisfied the three-step feature selection strategy in DC1.

| Variable | Step 1 | Step 2 | Step 3 | RFE |
|---|---|---|---|---|
| S Sensitivity_mean | Yes | Yes | Yes | Yes |
| Body check Table | Yes | No | - | - |
| Tea Towel First Response Indifferent | Yes | Yes | Yes | No |
| Distractions First Response Human | Yes | Yes | Yes | No |
| S Distractions Pull_mean | Yes | Yes | Yes | Yes |
| S Distractions Pull_prod | Yes | No | - | - |
| S Walking Distractions Pull_prod | Yes | No | - | - |
| S Crate Stimulus_mean | Yes | Yes | Yes | Yes |
| Crate Sniffing Exploring | Yes | No | - | - |
| Crate Actively Seeking Attention | Yes | Yes | Yes | No |
| S Petting Stimulus During_mean | Yes | Yes | Yes | Yes |
| S Petting Stimulus During_prod | Yes | No | - | - |
| S Sociability_mean | Yes | No | - | - |
| S Sociability_prod | Yes | No | - | - |
| Petting Handler Stimulus | Yes | No | - | - |
| Petting Confidence During | Yes | No | - | - |

Step 1 removed features that had $p > 0.15$ in the univariate logistic regression analysis as shown in Tables C.7, C.8, and C.9. It selected 15 out of the 111 variables; specifically, they were seven original items and eight composite scores. In Step 2, features that were associated with each other were removed. Only one variable was selected per group, resulting in the selection of three items and four scores. Finally, Step 3 considered the correlation between features to prevent multicollinearity. Figure 4.4 displays the correlation matrix: no variables were removed as none met the criteria.

Hence, seven features were inputted to the backward feature selection algorithm, which was set to maximise f1-score using LOOCV to estimate performance metrics. The optimal multivariate model created contained only composite scores; namely, “S Sensitivity_mean”, “S Distractions Pull_mean”, “S Crate Stimulus_mean”, and “S Petting Stimulus During_mean”. Table 4.3 shows the model parameters for the four covariates.

Except for “S Crate Stimulus_mean”, all of them had an estimate smaller than zero, indicating that an increase in the covariate decreases the probability of failure. This direction of association was also observed

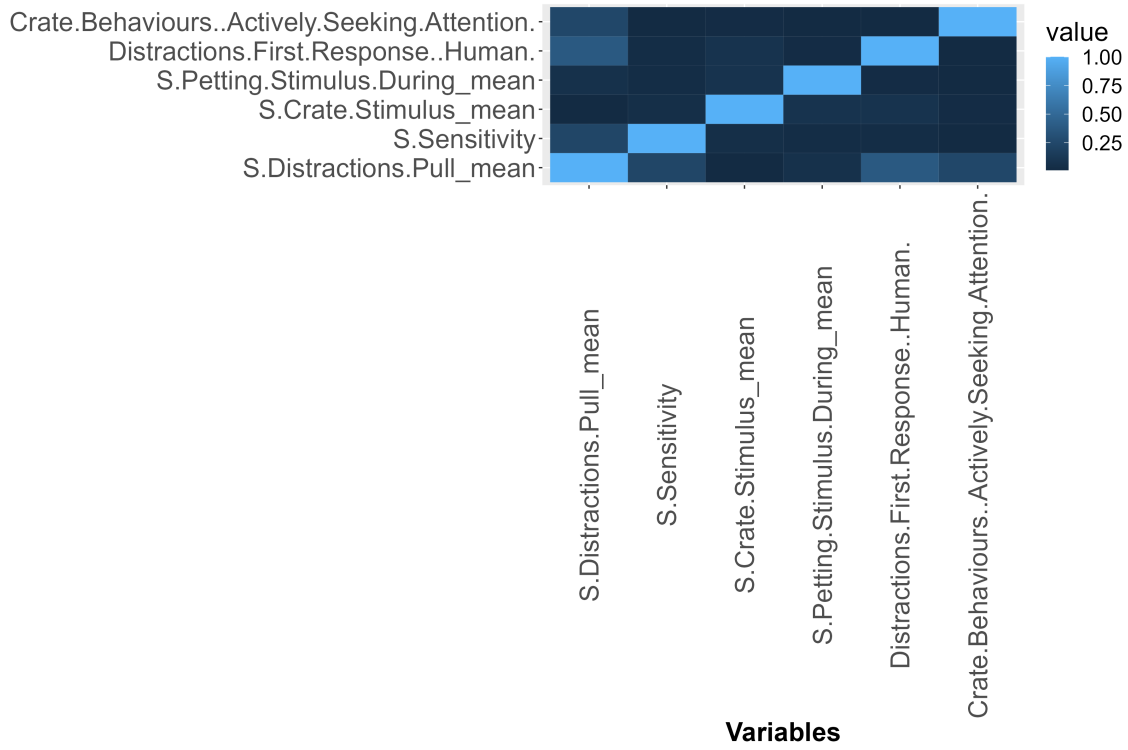



Figure 4.4: Correlation matrix for the numerical variables in Step 3 of feature selection.

Table 4.3: Optimal multivariate logistic regression model selected by RFE.

| Variable | Estimate | P-value | OR | 2.5% | 97.5% |
|---|---|---|---|---|---|
| Intercept **** | -1.67 | 0.00 | 0.19 | 0.09 | 0.34 |
| S Distractions Pull_mean**** | -0.91 | 0.00 | 0.40 | 0.21 | 0.72 |
| S Sensitivity_mean *** | -0.57 | 0.05 | 0.57 | 0.31 | 0.97 |
| S Crate Stimulus_mean * | 0.41 | 0.12 | 1.51 | 0.88 | 2.57 |
| S Petting Stimulus During_mean *** | -0.77 | 0.02 | 0.46 | 0.24 | 0.84 |

95% Confidence Interval (CI): lower 2.5% and upper 97.5%
Statistically significant with p < 0.01 (****), p < 0.05 (***), p < 0.10 (**), p < 0.15 (*)

from the univariate logistic regression models as indicated in Table C.7 and repeated in Table 4.4 for convenience.

Table 4.4: Univariate logistic regression model results of the variables in the optimal multivariate logistic regression model.

| Variable | Estimate | p-value | OR | 2.5% | 97.5% |
|---|---|---|---|---|---|
| S Sensitivity_mean ** | -0.83 | 0.06 | 0.43 | 0.17 | 1.03 |
| S Distractions Pull_mean *** | -0.58 | 0.05 | 0.56 | 0.30 | 0.99 |
| S Crate Stimulus_mean ** | 0.97 | 0.08 | 2.65 | 0.86 | 8.37 |
| S Petting Stimulus During_mean *** | -1.31 | 0.02 | 0.27 | 0.08 | 0.75 |

95% Confidence Interval (CI): lower 2.5% and upper 97.5%
Statistically significant with p < 0.05 (***), p < 0.10 (**), p < 0.15 (*)

These results can be interpreted to understand which types of behaviour contributed to a higher chance of failure. In order to do that, the original ethogram items used to calculate the composite scores should be identified in Table C.2, then the ethogram item scores in Table C.1. Table 4.5 combines key information from both tables for ease of reference.

The composite score “S Sensitivity_mean” was calculated by taking the average of the scores associated with the items “Body check Table” and “Body check Response”. High values of “S Sensitivity_mean” indicate that the dog had a high level of willingness to jump on the table and was comfortable with the body check procedure. The OR associated with this score was smaller than one, implying that high values of “S Sensitivity_mean” decrease the probability of failure. Accordingly, the model learnt that dogs scoring high on willingness to jump on the table and comfort with the body check procedure had a lower probability of failing.

The composite score “S Distractions Pull_mean” was calculated by taking the average of the scores associated with the items “Pull on leash” and “Pull strength”. High values of “S Distractions Pull_mean” are produced when the dog pulls on the leash frequently and/or strongly. The OR associated with this score was smaller than one, implying that high values of “S Distractions Pull_mean” decrease the probability of failure. Hence, the model learnt that dogs scoring high on frequency and strength of pull have a lower probability of failing.

The composite score “S Crate Stimulus_mean” was calculated by taking the average of the frequency of behaviours directed towards the crate, in particular: “Sniffing/Exploring”, “Digging”, “Nudging Crate”, and “Pawing”. High values of “S Crate Stimulus_mean” indicate that the dog showed a high level of engagement with the crate. The OR associated

Table 4.5: Composite scores, original ethogram items, and their respective marking for the variables chosen for the optimal model.

| Subtest | Items | Marking |
|---|---|---|
| **S Sensitivity_mean** | | |
| 6. Body check | Table | 1 - Refuse<br>2 - Hesitant<br>3 - Cooperative<br>4 - Willing |
| | Response | 1 - Escaping/Not willing<br>2 - Reluctant<br>3 - Cooperative<br>4 - Willing |
| **S Distractions Pull_mean** | | |
| 8. Distractions | Pull on leash | 1 - Not observed (0)<br>2 - Rare (1-2)<br>3 - Sometimes (3-5)<br>4 - Frequent (6-8) |
| | Pull strength | 1 - No tension, lead relaxed<br>2 - Mild tension<br>3 - Moderate tension<br>4 - Strong tension<br>5 - Extreme tension, back legs stretched, and one or more front paws raised off the floor |
| **S Crate Stimulus_mean** | | |
| 11. Crate | Sniffing/Exploring<br>Digging<br>Nudging Crate<br>Pawing | 1 - Never<br>2 - Rarely<br>3 - Occasionally<br>4 - Regularly<br>5 - Constantly |
| **S Petting Stimulus During_mean** | | |
| 12. Petting | Handler-Stimulus | 1 - Slightly playful<br>2 - Moderately playful<br>3 - Extremely playful |
| | Confidence-During | 1 - Avoids, moves away<br>2 - Indifferent<br>3 - Slightly excited, tail wagging<br>4 - Very excited, moving around<br>5 - Extremely excited, jumping up, licking, mouthing |

with this score was bigger than one, implying that high values of "Petting Handler-Stimulus" increase the probability of failure. Therefore, the model learnt that dogs that frequently engage with the crate have a higher probability of failing.

The composite score "S Petting Stimulus During" was calculated by taking the average of the scores associated with the items "Petting Handler-Stimulus" and "Petting Confidence-During". High values of "S Petting Stimulus During" indicate that the dog showed a high level of excitement when playing with their handler and likewise their handler was quite playful towards the dog. The OR associated with this score was smaller than one, implying that high values of "Petting Handler-Stimulus" decrease the probability of failure. Thus, the model learnt that dogs scoring high on excitement towards their handler during play have a lower probability of failing.

Table C.6 in Appendix C.3.1 shows the descriptive statistics for the variables in the optimal multivariate logistic regression model, and the items from which they were calculated. All items used the full range except for three variables from subtest 11 "crate", namely, "Crate Digging", "Crate Nudging Crate", and "Crate Pawing", as indicated in Table 4.5.

Model fitness was evaluated considering the reduction in deviance resulting from the addition of the four variables. The null deviance was 101.84 on 97 degrees of freedom and the residual deviance was 81.84 on 93 degrees of freedom. Hence, the difference in deviance was $101.84 - 81.84 = 20.00$ on $97 - 93 = 4$ degrees of freedom. The model can be considered a good fit for the data as it fits significantly better than the null model $P(\tilde{\chi}_4^2 \geq 20.00) = 5.00e^{-04}$.

All performance metrics were estimated from the predictions made by the optimal model, with four composite scores derived from the ethogram items using LOOCV. The ROC and PR curves are shown in Figure 4.5. The optimisation metric f1-score was 0.32; the area under the ROC curve was 0.74; and the PR curve was 0.37. Precision, specificity, and sensitivity were 0.50, 0.94, and 0.24, respectively.

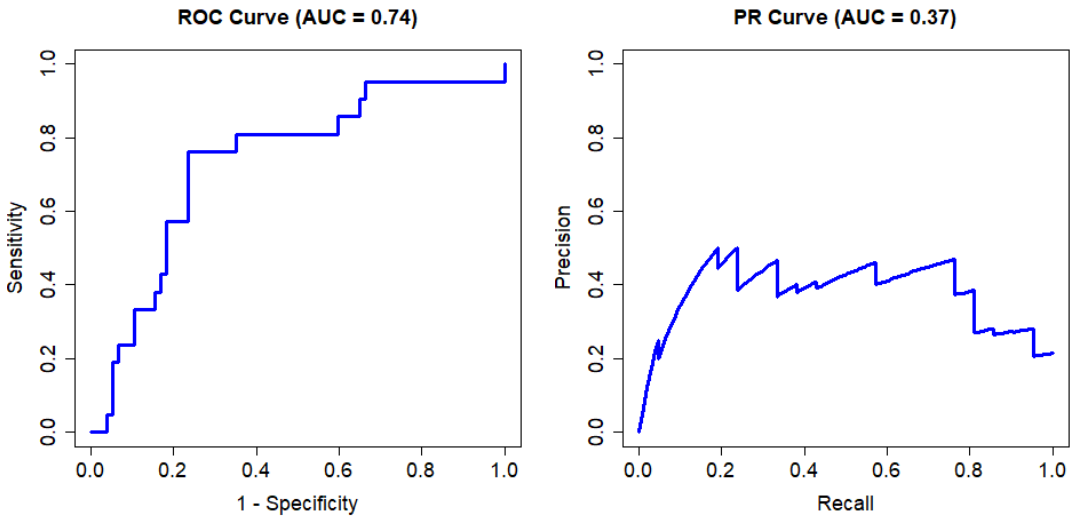



Figure 4.5: ROC and PR curves for the optimal logistic regression model using ADTB data from DC1 selected by RFE.

Figures 4.6 and 4.7 show the probability of failure predicted by the best logistic regression model for all the 98 dogs in the dataset. In particular, Figure 4.6 displays the working outcome obtained by each dog, i.e., whether the dog failed or succeed at graduating and working for one year. Figure 4.7 shows the same probability curve highlighting whether the dog became a guide dog, or assistance dog, or was withdrawn from training. It can be visually observed that guide dogs were assigned the lowest probabilities of failure and withdrawn dogs had the highest probabilities.

The lowest probability of failure belonged to guide dogs (M = 0.14, SD = 0.13), assistance dogs (M = 0.24, SD = 0.24), and withdrawn dogs (M = 0.62, SD = 0.21). There was a statistically significant difference in the probability of failure for different training outcome statuses (H(2) = 14.04, p = 8.929e-04), in particular between guide dogs and withdrawn dogs (p = 2.7e-04).

Table 4.6 shows the results of the feature importance analysis of the optimal multivariate logistic regression model. In order of importance, they were “S Distractions Pull_mean”, “S Petting Stimulus During_mean”, “S Sensitivity_mean”, and “S Crate Stimulus_mean”.

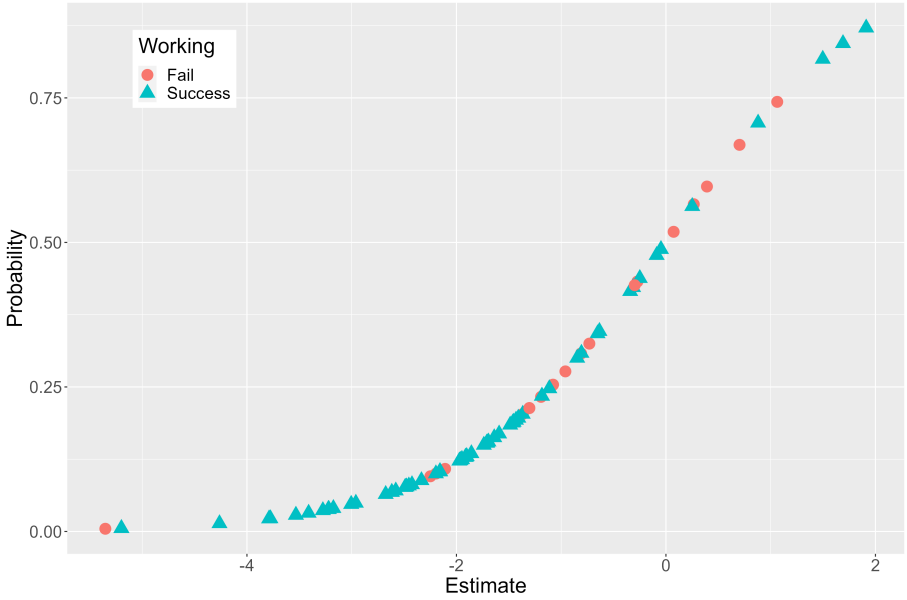



Figure 4.6: Probability of failure predicted by the best logistic regression model using DC1 data per true working outcome (fail, success).

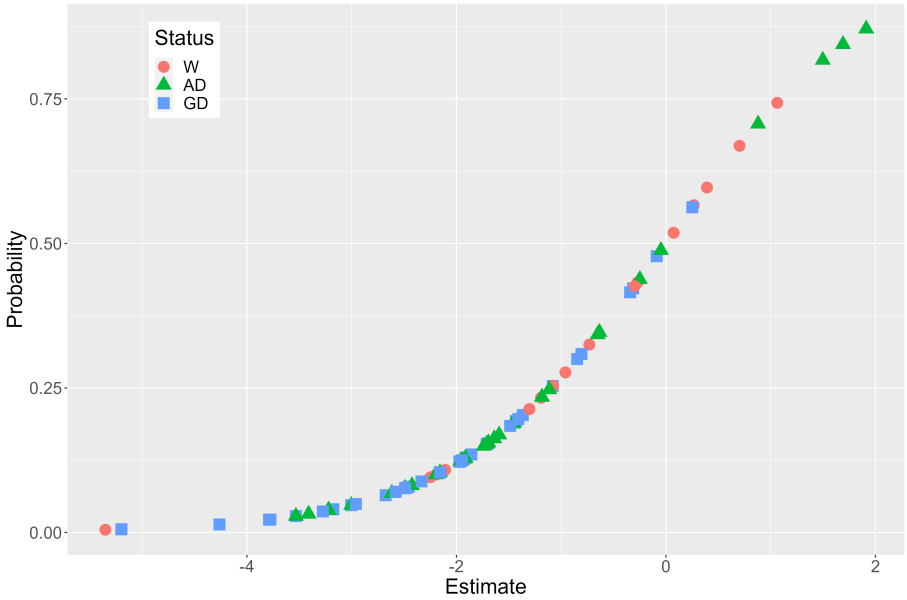



Figure 4.7: Probability of failure predicted by the best logistic regression model using DC1 data per true status at the end of training (W = withdrawn, AD = assistance dog, and GD = guide dog).

### 4.3.2 Data Collection 2

The sample comprised a total of 60 dogs, including 32 females (53%) and 28 males (47%); there were 25 Labrador Retriever, 34 Labrador Retriever crossed with Golden Retriever, and 1 Golden Retriever. Their age at arrival ranged from 13 to 22 months (M = 15.3, SD = 1.9). Their age at Data Collection 2 varied from 15 to 24 months (M = 17.8, SD = 1.8). The

Table 4.6: Importance of features in the optimal model.

| Variable | Importance |
|---|---|
| S Distractions Pull_mean | 2.89 |
| S Petting Stimulus During_mean | 2.40 |
| S Sensitivity_mean | 2.00 |
| S Crate Stimulus_mean | 1.54 |

failure rate in this sample was 20%, as 12 dogs failed and 48 dogs succeeded in their training.

Figure 4.8 shows the duration of training for the dogs withdrawn from training for behavioural reasons. Dogs were withdrawn at around 18 weeks of training on average. Only one dog had been withdrawn before 10 weeks. Only one dog was withdrawn during the first year of the working programme.

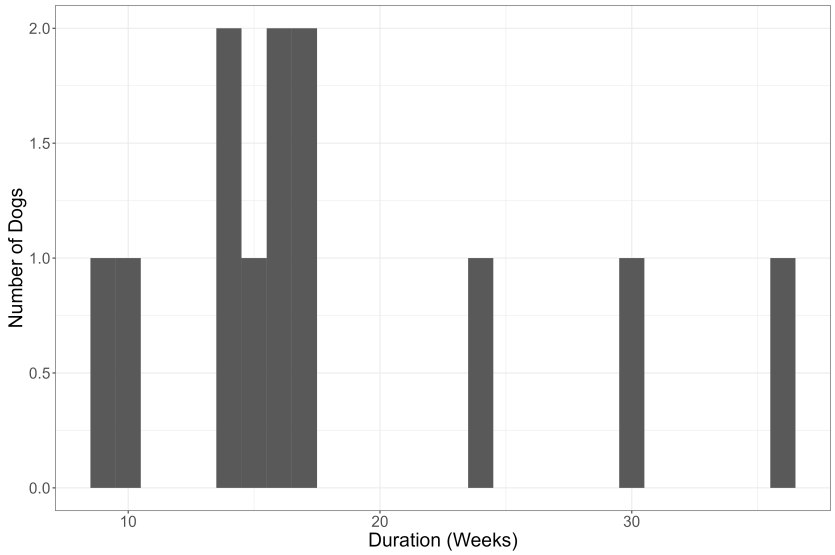



Figure 4.8: Duration of training in weeks for dogs that were withdrawn from training.

Table C.10 shows descriptive statistics of the numerical variables in this sample including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing data of all original items and composite scores. Table C.11 describes the number of instances per category and missing data in this sample. The marking scale used to score the items is described in Table C.1.

#### 4.3.2.1 Validity

Univariate logistic regression was used to investigate whether there was a statistically significant difference between the working outcome of the dogs and ADTB ethogram items and scores, and their demographic

information. The results from the logistic regression univariate models are shown in Table C.14, highlighting three levels of significance. A total of 11 variables had p < 0.05, six were in the range 0.05 < p < 0.10 and seven were within 0.10 < p < 0.15. The categorical variables "Breed" and "Kong Interaction Lateralisation" had more than two categories, thus their results are shown in Tables C.15 and C.16.

Variables with a negative estimate decrease the probability of failure as their scores increase. Conversely, variables with a positive estimate increase the probability of failure as their scores increase. Seven out of the 11 variables that showed a significant association with working outcomes at traditional levels (p < 0.05) had a negative estimate, in ascending order: "Tea Towel Second Response Indifferent", "Walking Pull strength", "S Walking Pull_mean", "Walking Pull on leash", "S Isolation Handler_mean", "Isolation Response Time oriented", and "S Walking Pull_prod". The other four variables had a positive estimate: "Isolation Response Exploration", "S Isolation Stimulus", "Distractions First Response Car", and "Kong Interaction Response to stimulus".

#### 4.3.2.2 Predictability

A three-step feature selection strategy was carried out. The initial dataset contained two demographic variables; 70 original items and 41 composite scores. Table 4.7 indicates which of the variables selected in step 1 satisfied steps 2 and 3, and were selected by the RFE algorithm. They were arranged according to the order of the subtests as shown in Table 4.1 and the lines separate groups of variables that are related to each other as indicated in Table C.2.

Step 1 removed features that had p > 0.15 in the univariate logistic regression analysis as shown in Tables C.14, C.15 and C.16. It selected 23 out of the 111 ADTB variables; specifically, they were 12 original items and 11 composite scores. In Step 2, features that were associated with each other were removed. Only one variable was selected per group, resulting in the selection of two items and seven scores. Finally, Step 3 considered the correlation between features to prevent multicollinearity. Figure 4.9 displays the correlation matrix for numerical variables, the categorical variable "Tea Towel Second Response Change from Neutral" was removed because it was significantly related to "Tea Towel Second Response Indifferent" ($P(\tilde{\chi}_1^2 \geq 36.797) = 1.31e^{-09}$.

Hence, nine features were input to the backward feature selection algorithm which was set to maximise f1-score using LOOCV to estimate performance metrics. The optimal multivariate model created contained one ethogram item and five composite scores; namely, "Distractions First Response Car", "S Walking Pull_mean", "S Familiarisation Handler_mean", "S Isolation Stimulus", "S Crate

Table 4.7: Variables that satisfied the three-step feature selection strategy in DC2.

| Variable | Step 1 | Step 2 | Step 3 | RFE |
|---|---|---|---|---|
| S Familiarisation Handler_mean | Yes | Yes | Yes | Yes |
| S Familiarisation Handler_prod | Yes | No | - | - |
| S Walking Pull_mean | Yes | Yes | Yes | Yes |
| S Walking Pull_prod | Yes | No | - | - |
| S Walking Distractions Pull_mean | Yes | No | - | - |
| Walking Pull on leash | Yes | No | - | - |
| Walking Pull strength | Yes | No | - | - |
| S Sensitivity_mean | Yes | Yes | Yes | Yes |
| Body check Table | Yes | No | - | - |
| Tea Towel Second Response Indifferent | Yes | Yes | Yes | No |
| Tea Towel Second Response Change from Neutral | Yes | Yes | No | - |
| Distractions First Response Car | Yes | Yes | Yes | Yes |
| S Kong Response_prod | Yes | Yes | Yes | No |
| Kong Interaction Response to stimulus | Yes | No | - | - |
| Kong Return Handler | Yes | No | - | - |
| S Crate Handler_mean | Yes | Yes | Yes | Yes |
| S Crate Handler_prod | Yes | No | - | - |
| Crate Actively Seeking Attention | Yes | No | - | - |
| Crate Whining | Yes | No | - | - |
| S Isolation Handler_mean | Yes | Yes | Yes | No |
| Isolation Response Time oriented | Yes | No | - | - |
| S Isolation Stimulus | Yes | Yes | Yes | Yes |
| Isolation Response Exploration | Yes | No | - | - |

Handler_mean", and "S Sensitivity". Table 4.8 shows the model parameters for the six covariates.

Except for "Distractions First Response Car" and "S Isolation Stimulus", all of the variables had an estimate smaller than zero. The same direction of association was also observed from the univariate logistic regression models as indicated in Table C.14 and repeated in Table 4.9 for convenience.

Quantitative analysis of the model is not possible due to the data pre-processing techniques employed (normalisation). The model estimates can only be interpreted qualitatively to understand which types of behaviour contributed to a higher chance of failure. In order to interpret estimates associated with composite scores, the original ethogram items used to calculate the composite scores were identified in Table C.2 following the marking system in Table C.1. Table 4.10

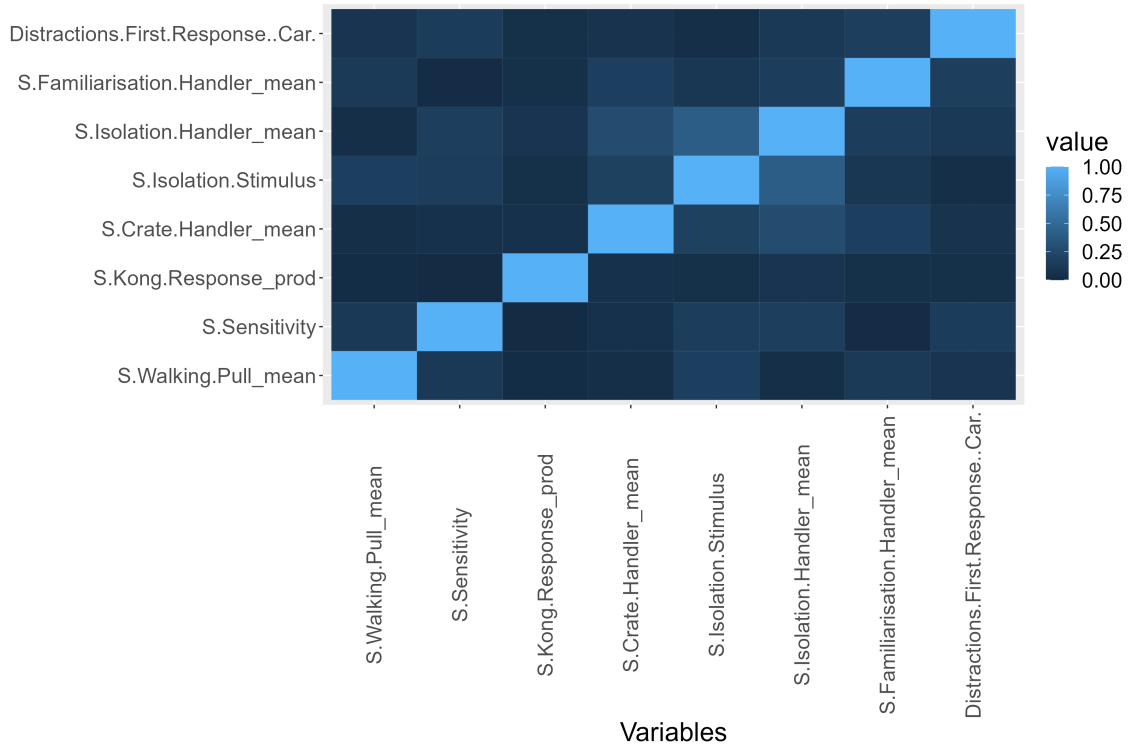



Figure 4.9: Correlation matrix for the numerical variables in Step 3 of feature selection.

Table 4.8: Optimal multivariate logistic regression model using selected by RFE.

| Variable | Estimate | P-value | OR | 2.5% | 97.5% |
|---|---|---|---|---|---|
| Intercept **** | -3.26 | 0.00 | 0.04 | 0.00 | 0.17 |
| Distractions First Response Car *** | 1.43 | 0.02 | 4.18 | 1.39 | 18.16 |
| S Walking Pull_mean *** | -1.88 | 0.02 | 0.15 | 0.02 | 0.57 |
| S Familiarisation Handler_mean ** | -1.59 | 0.08 | 0.20 | 0.02 | 0.77 |
| S Isolation Stimulus | 0.66 | 0.23 | 1.93 | 0.71 | 6.79 |
| S Crate Handler_mean | -1.13 | 0.26 | 0.32 | 0.03 | 1.75 |
| S Sensitivity_mean | -0.48 | 0.27 | 0.62 | 0.25 | 1.49 |

95% Confidence Interval (CI): lower 2.5% and upper 97.5%
Statistically significant with p <0.01 (****), p < 0.05 (***), p < 0.10 (**), p < 0.15 (*)

Table 4.9: Univariate logistic regression model results of the variables in the optimal multivariate logistic regression model.

| Variable | Estimate | p-value | OR | 2.5% | 97.5% |
|---|---|---|---|---|---|
| Distractions First Response Car *** | 0.79 | 0.02 | 2.21 | 1.18 | 4.71 |
| S Familiarisation Handler_mean * | -0.96 | 0.10 | 0.38 | 0.10 | 1.01 |
| S Walking Pull_mean *** | -1.72 | 0.02 | 0.18 | 0.03 | 0.63 |
| S Sensitivity_mean ** | -0.86 | 0.09 | 0.42 | 0.15 | 1.14 |
| S Crate Handler_mean ** | -1.25 | 0.09 | 0.29 | 0.04 | 0.92 |
| S Isolation Stimulus *** | 0.60 | 0.03 | 1.83 | 1.10 | 3.25 |

95% Confidence Interval (CI): lower 2.5% and upper 97.5%
Statistically significant with $p < 0.05$ (***), $p < 0.10$ (**), $p < 0.15$ (*)

combines key information from both tables for ease of reference.

High values of the ethogram item “Distractions First Response Car” were assigned to dogs who got extremely distracted by the moving car during the Distraction subtest. The OR associated with this variable was bigger than one, indicating that high values of “Distractions First Response Car” increased the probability of failure. Thus, the model learnt that dogs who were more distracted in subtest 8 “Distractions” by the remote-controlled car had a higher probability of failing.

The composite score “S Familiarisation Handler_mean” was calculated by taking the average of the scores associated with the items “Familiarisation Oriented to Handler”, and “Familiarisation Waiting”. High values of “S Familiarisation Handler_mean” indicate that the dog was constantly oriented to the handler or waiting during the Familiarisation subtest. The OR associated with this score was smaller than one, implying that high values of “S Familiarisation Handler_mean” decreased the probability of failure. Accordingly, the model learnt that dogs who spent longer oriented to the handler or waiting during subtest 1 “Familiarisation” had a lower probability of failing.

The composite score “S Walking Pull_mean” was calculated by taking the average of the scores associated with the items “Walking Pull on leash” and “Walking Pull strength”. High values of “S Walking Pull_mean” indicate that the dog was constantly and strongly pulling on the leash during the Walking subtest. The OR associated with this score was smaller than one, implying that high values of “S Walking Pull_mean” decreased the probability of failure. Hence, the model learnt that dogs who constantly and strongly pulled on their leashes during subtest 2 “Walking” had a lower probability of failing.

The composite score “S Sensitivity_mean” was calculated by taking the

Table 4.10: Composite scores, original ethogram items, and their respective marking for the variables chosen for the optimal model.

| Subtest | Items | Marking |
|---|---|---|
| **S Familiarisation Handler_mean** | | |
| 1. Familiarisation | Oriented to Handler; Waiting | 1 - Never; 2 - Rarely; 3 - Occasionally; 4 - Regularly; 5 - Constantly |
| **S Walking Pull_mean** | | |
| 2. Walking | Pull on leash | 1 - Not observed (0) ; 2 - Rare (1-2); 3 - Sometimes (3-5); 4 - Frequent (6-8) |
| | Pull strength | 1 - No tension, lead relaxed; 2 - Mild tension; 3 - Moderate tension; 4 - Strong tension; 5 - Extreme tension, back legs stretched, and on or more front paws raised off the floor |
| **S Sensitivity_mean** | | |
| 6. Body check | Table | 1 - Refuse 2 - Hesitant 3 - Cooperative 4 - Willing |
| | Response | 1 - Escaping/Not willing 2 - Reluctant 3 - Cooperative 4 - Willing |
| **S Crate Handler_mean** | | |
| 11. Crate | Actively Seeking Attention; Whining; Barking | 1 - Never; 2 - Rarely; 3 - Occasionally; 4 - Regularly; 5 - Constantly |
| **S Isolation Stimulus** | | |
| 13. Isolation | Exploration; Unsettled/Pacing | 1 - Never; 2 - Rarely; 3 - Occasionally; 4 - Regularly; 5 - Constantly |

average of the scores associated with the items “Body check Table” and “Body check Response”. High values of “S Sensitivity_mean” indicate that the dog had a high level of willingness to jump on the table and/or comfort with the body check procedure. The OR associated with this score was smaller than one, implying that high values of “S Sensitivity_mean” decreased the probability of failure. Accordingly, the model learnt that dogs who scored high on willingness to jump on the table and comfort with touch during subtest 6 “Body check” had a lower probability of failing.

The composite score “S Crate Handler_mean” was calculated by taking the average of the scores associated with the items “Crate Actively Seeking Attention”, “Crate Whining”, and “Crate Barking”. High values of “S Crate Handler_mean” indicate that the dog more frequently tried to get their handler’s attention with general behaviours, whining, and/or barking. The OR associated with this score was smaller than one, implying that high values of “S Crate Handler_mean” decreased the probability of failure. Therefore, the model learnt that dogs who scored high on attention-seeking behaviours during subtest 11 “Crate” had a lower probability of failing.

Finally, the composite score “S Isolation Stimulus” was calculated by taking the average of the scores associated with the items “Exploration” and “Unsettled/Pacing”. High values of “S Isolation Stimulus” indicate that the dog mainly moved around the test arena while their handler was absent. The OR associated with this score was higher than one, implying that high values of “S Isolation Stimulus” increase the probability of failure. Thus, the model learnt that dogs that explored more frequently and paced less frequently during subtest 13 “isolation” had a higher probability of failing.

Table C.13 in Appendix C.4.1 shows the descriptive statistics for the variables in the optimal multivariate logistic regression model and the items from which they were calculated. The only ethogram items that did not reach the whole range were “Walking Pull strength” and “Isolation Response Unsettled/Pacing”. Considering all categories shown in Table 4.10, two categories of behaviour were not observed in the former (“4 - Strong tension” and “5 - Extreme tension”) and one category in the latter (“5 - Constantly”) during DC2.

Model fitness was evaluated considering the reduction in deviance resulting from the addition of the six variables. The null deviance was 60.05 on 59 degrees of freedom and the residual deviance was 29.06 on 53 degrees of freedom. Hence, the difference in deviance was $60.05 - 29.06 = 30.99$ on $59 - 53 = 6$ degrees of freedom. The model can be considered a good fit for the data as it fits significantly better than the null model $P(\tilde{\chi}^2_6 \geq 30.99) = 2.54e^{-05}$.

All performance metrics were estimated from the predictions made by

the optimal model with one ethogram item and five composite scores using LOOCV. The ROC and PR curves are shown in Figure 4.10. The optimisation metric f1-score was 0.67, the area under the ROC curve was 0.84, and the PR curve was 0.74. Precision, specificity, and sensitivity were 0.67, 0.92, and 0.67, respectively.

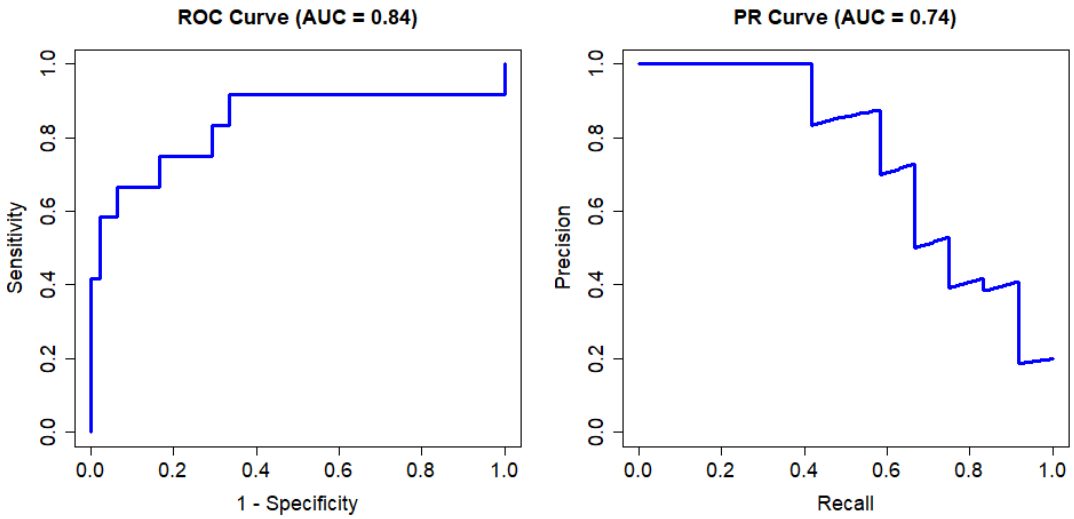



Figure 4.10: ROC and PR curves for the optimal logistic regression model using ADTB data from DC2 selected by RFE.

Figures 4.11 and 4.12 show the probability of failure predicted by the best logistic regression model for all the 60 dogs in the dataset. In particular, Figure 4.11 displays the working outcome obtained by each dog, i.e., whether the dog failed or succeeded at graduating and working for one year. Figure 4.12 shows the same probability curve highlighting whether the dog became a guide dog, assistance dog, or was withdrawn from training. It can be visually observed that guide dogs were assigned the lowest probabilities of failure and withdrawn dogs had the highest probabilities.

The lowest probability of failure belonged to guide dogs (M = 0.04, SD = 0.10), assistance dogs (M = 0.22, SD = 0.25), and withdrawn dogs (M = 0.62, SD = 0.33). There was a statistically significant difference in the probability of failure for different training outcome statuses (H(2) = 26.2, p = 2.005e-06) between all groups. The differences between guide dogs and withdrawn dogs (p = 7.5e-08), guide dogs and assistance dogs (p = 6.9e-04), and assistance and withdrawn dogs (p = 2.2e-03) were highly significant.

Table 4.11 shows the results of the feature importance analysis of the optimal multivariate logistic regression model. In order of importance, they were “Distractions First Response Car”, “S Walking Pull_mean”, “S

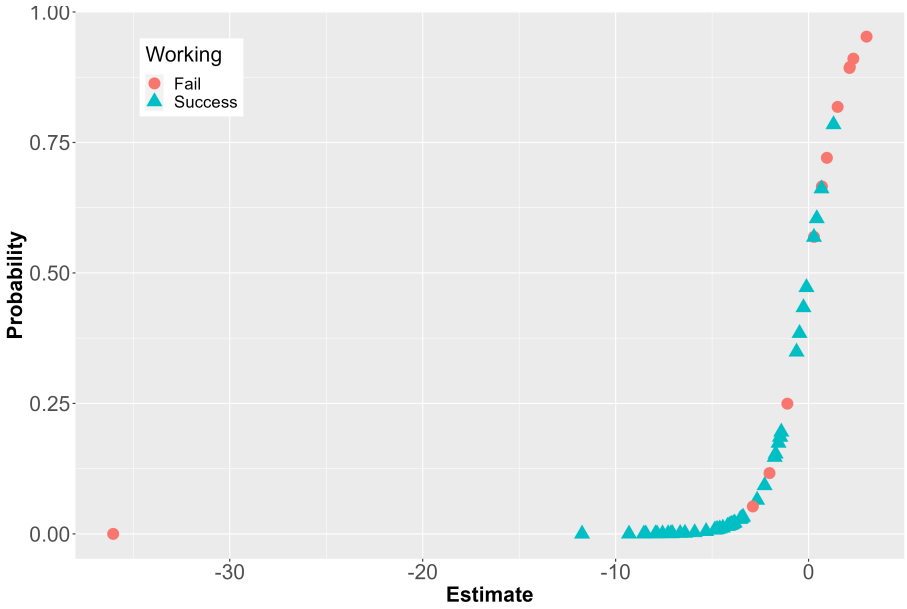



Figure 4.11: Probability of failure predicted by the best logistic regression model using DC2 data per true working outcome (fail, success).

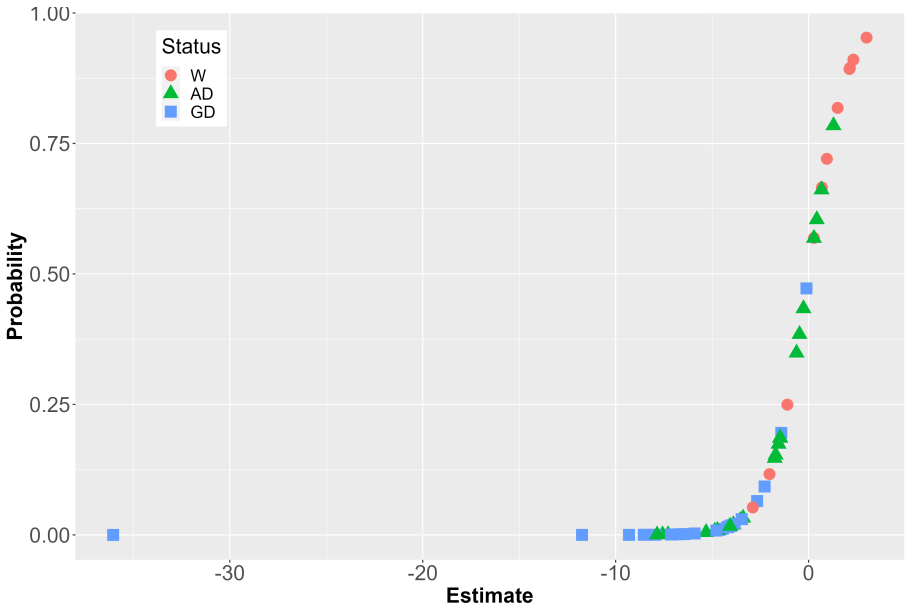



Figure 4.12: Probability of failure predicted by the best logistic regression model using DC2 data per true status at the end of training (W = withdrawn, AD = assistance dog, and GD = guide dog).

Familiarisation Handler_mean", "S Isolation Stimulus", "S Crate Handler_mean", and "S Sensitivity_mean".

Table 4.11: Importance of features in the optimal model.

| Variable | Importance |
|---|---|
| Distractions First Response Car | 2.27 |
| S Walking Pull_mean | 2.30 |
| S Familiarisation Handler_mean | 1.76 |
| S Isolation Stimulus | 1.19 |
| 1 S Crate Handler_mean | 1.13 |
| S Sensitivity_mean | 1.10 |

## 4.4 Discussion

In this study, machine learning techniques were successfully employed to predict future working outcomes in assistance dogs using the novel ADTB. Working outcomes were more definitive measure of true success than training outcomes (L. Batt et al. 2010). Previous research mainly focused on predicting graduation/training outcomes, except for one study (Bray et al. 2019). The results demonstrate the superior performance of the new ADTB over the IFT, as the new models were capable of more reliably distinguishing between the two working outcomes in comparison to the state-of-the-art.

In particular, the DC1 and DC2 models achieved an ROC of 0.74 and 0.84, respectively, which is higher than the 0.71 presented in Bray et al. 2019. The DC2 model constitutes a substantial improvement over previously reported research, while the DC1 model provided a moderate enhancement. Accordingly, this study contributes to the body of knowledge, by providing a new method that more effectively predicts working outcomes in apprentice assistance dogs. Moreover, in addition to the behavioural assessment dataset, inertial, and physiological datasets monitored during the test battery will also be made publicly available.

The optimal models built using DC1 and DC2 to predict the probability of failure are analysed in Sections 4.4.1 and 4.4.2, respectively. The validity of the models was analytically assessed using the cross-validated prediction performance metrics and visually observed using Figures 4.6 and 4.11. Additionally, working outcome predictions were categorised by outcome status in Figures 4.7 and 4.12. Even though the model predicted the probability of future working dog failure, the plot reveals that guide dogs were predicted to have the lowest likelihood of failure, followed by assistance dogs, and finally withdrawn dogs had the highest chance of failure.

The relationship between independent variables in the optimal model and future working dog outcomes can be assessed based on the sign (positive or negative) of the estimate. A positive sign indicates that high

values increase the probability of failure, and a negative sign implies that higher scores decrease the probability of failure. The independent variables originating from the ethogram items are more straightforward to interpret than those from composite scores, as the former is calculated based on a combination of ethogram items.

Section 4.4.3 compares results from DC1 and DC2, including the relationship between working outcome with ethogram items and composite scores, as well as parameters and performance metrics of the optimal models. Section 4.4.4 discusses the results achieved in this study and previously published research.

### 4.4.1 Data Collection 1

The optimal model was built using four composite scores recorded during four different subtests; namely, “body check”, “distractions”, “crate”, and “ petting”. The predicted probability of failure increased as the value of “S Crate Stimulus_mean” increased. Thus, a higher likelihood of failure was estimated for dogs who more frequently exhibited behaviours towards the crate such as sniffing, digging, nudging, and pawing in the subtest 11 “crate”.

On the other hand, the predicted probability of failure decreased as the value of the following variables increased: “S Sensitivity_mean”, “S Distractions Pull_mean”, and “S Petting Stimulus During_mean”. A lower likelihood of failure was assigned to dogs who pulled on the leash frequently and/or strongly in subtest 8 “distractions”; were willing to jump on the table and/or comfortable with touch in subtest 6 “body check”; and were highly engaged with their handler during subtest 12“petting”.

### 4.4.2 Data Collection 2

The optimal model was built using variables recorded during six subtests; namely, “familiarisation”, “walking”, “body check”, “distractions”, “crate”, and “isolation”. The predicted probability of failure increased as the value of the following variables increased: “Distractions First Response Car” and “S Isolation Stimulus”. Accordingly, a higher likelihood of failure was estimated by a high level of distraction caused by the moving car in the first instance of subtest 8 “distractions”; and a high value of “S Isolation Stimulus” which was created by a constant exploration and/or pacing around the test arena in subtest 13 “isolation”.

The probability of failure decreased as the value of the following scores increased: “S Familiarisation Handler_mean”, “S Walking Pull_mean”, “S Sensitivity_mean”, and “S Crate Handler_mean”. A lower likelihood

of failure was assigned to dogs who spent longer oriented to the handler or waiting in subtest 1 “familiarisation”; pulled on the leash constantly and/or strongly in subtest 2 “walking”; and were willing to jump on the table and/or were comfortable with touch in subtest 6 “body check”; and were frequently actively seeking attention, whining or barking in the subtest 11 “crate”.

### 4.4.3 Comparison between DC1 and DC2 Models

Some data collection sessions had to be cancelled due to the global COVID-19 pandemic resulting in a reduced number of samples in DC2 (N = 60) compared to DC1 (N = 98). In terms of timing, assistance dog training organisations are particularly interested in performing a behaviour test for the prediction of working outcomes at the earliest possible stage of training. Hence, it is preferable to use the model built using DC1 instead of DC2 in case their performance is comparable.

Step 1 in the feature selection strategy involved the evaluation of the association between each variable and working outcome, using univariate logistic regression. Table 4.12 shows the 15 variables selected in DC1 and 23 in DC2.

Table 4.12: Ethogram items and composite scores are separated per subtest and by Data Collection 1 (DC1) and Data Collection 2 (DC2). Variables with a level of significance from univariate logistic regression below the threshold for retention ($p < 0.15$) are included. The variables highlighted in bold were selected for the optimal multivariate models.

| Subtest | DC1 | DC2 |
|---|---|---|
| 1. Familiarisation | | **S Familiarisation Handler_mean** *<br>S Familiarisation Handler_prod * |
| 2. Walking | | **S Walking Pull_mean** ***<br>S Walking Pull_prod ***<br>S Walking Distractions Pull_mean **<br>Walking Pull on leash ***<br>Walking Pull strength *** |
| 3. Standing | | |
| 4. Sitting | | |
| 5. Lying down | | |
| 6. Body check | Table **<br>**S Sensitivity_mean** ** | Table *<br>**S Sensitivity_mean** ** |
| 7. Tea towel | First Response Indifferent * | Second Response Indifferent **<br>Second Response Change from Neutral * |
| 8. Distractions | First Response Human *<br>**S Distractions Pull_mean** **<br>S Distractions Pull_prod ** | **First Response Car** *** |
| 9. Kong | | Interaction Response to stimulus ***<br>Return Handler **<br>S Kong Response_prod * |
| 10. Dog distraction | | |
| 11. Crate | Actively Seeking Attention *<br>Sniffing Exploring ** | Actively Seeking Attention **<br>Whining * |

| | | |
|---|---|---|
| | **S Crate Stimulus_mean** ** | **S Crate Handler_mean** **<br>S Crate Handler_prod * |
| 12. Petting | Handler Stimulus *<br>Confidence During ***<br>**S Petting Stimulus During_mean** **<br>S Petting Stimulus During_prod **<br>S Sociability_mean **<br>S Sociability_prod * | |
| 13. Isolation | | Response Time oriented ***<br>Response Exploration ***<br>**S Isolation Stimulus** ***<br>S Isolation Handler_mean *** |
| 14. Noise | | |

Statistically significant with $p < 0.05$ (***), $p < 0.10$ (**), $p < 0.15$ (*)

Variables associated with the "body check", "tea towel", "distractions", and "crate" subtests met the threshold for retention in both DC1 and DC2. In particular, "Table" in subtest 6 "body check" and "Actively Seeking Attention" in subtest 11 "crate" were the only two ethogram items, and "S Sensitivity_mean" was the only composite score retained in step 1 of both data collections. A similar variable was selected in subtest 8 "tea towel", namely "First Response Indifferent" in DC1 and "Second Response Indifferent" in DC2.

Variables from "petting" met the threshold for retention in DC1, but not in DC2. Variables from "familiarisation", "walking", "Kong", and "isolation" subtests met the threshold for retention in DC2, but not in DC1. No variables from the following five subtests achieved the threshold for retention in both DC1 and DC2: "standing", "sitting", "lying down", "dog distraction", and "noise". Thus, the data indicate that these subtests were not particularly relevant for the prediction of working outcomes.

Conclusively, the data suggest that the most important subtests for the prediction of working outcome were: "body check", "distractions", and "crate" in both DC1 and DC2; "petting" in DC1 only; "familiarisation", "walking", "body check", "distractions", "crate", and "isolation" in DC2. Hence, considering that "body check", "distractions", and "crate" were the subtests that provided variables for both optimal models, they can be considered the most relevant across time.

Table 4.13 shows the performance metrics from the models created with data from DC1 and DC2. All performance metrics reported were significantly better for DC2 compared to DC1, except for specificity which showed a marginal drop. In particular, the biggest differences were noticed in sensitivity, the area under the PR curve, and f1-score as there was a significant improvement in the correct detection of true positives (failure).

Table 4.13: Performance metrics from the models created with data from Data Collection 1 (DC1) and Data Collection 2 (DC2).

| Dataset | ROC | PR | Precision | Sensitivity | Specificity | F1 |
|---|---|---|---|---|---|---|
| DC1 | 0.74 | 0.37 | 0.50 | 0.24 | 0.94 | 0.32 |
| DC2 | 0.84 | 0.74 | 0.67 | 0.67 | 0.92 | 0.67 |

These performance results indicate that the predictions computed by the DC2 model are significantly more reliable than those generated by the DC1 model. Although earlier predictions are preferable, it is fundamental for the predictions to be as accurate as possible. Hence, it is recommended to evaluate trainee dogs using the ADTB at 10 weeks after the start of their formal training, in order to estimate their suitability for assistance work using the DC2 model.

### 4.4.4 Comparison with Previous Research

Two types of comparison are discussed. In Section 4.4.4.1, the results from subtests included in this study are compared with previous research which included similar subtests aiming at understanding the relationship between specific behaviours shown by assistance dogs and future outcomes. In Section 4.4.4.2, performance metrics reported in research focused on predicting future outcomes in assistance dogs are compared with the present study. Different research studies report different metrics which are visually organised in a table and discussed.

#### 4.4.4.1 Test Batteries and Training Outcome

In addition to the assistance dog prediction model performance metrics, it is also of interest to discuss the relationship between individual subtests and future outcomes in previous research and this research. Hence, the remainder of this section fulfils that intent as each of the following paragraphs concerns one of the subtests.

Subtest 1 "familiarisation" was not analysed in terms of training outcomes in assistance dogs in previous research. The final predictive model in DC1 used the composite score "S Familiarisation Handler_mean", which combined the time-oriented to the handler and waiting for the handler's commands from this subtest.

Subtest 2 "walking" was similar to the subtest "leash" in Wilsson et al. 2012 during which dogs were rated based on the level of support they required from the handler. In that study, variables from this subtest contributed to the principal component most predictive of training outcomes in trainee military working dogs. The results revealed that a high level of dependence on cues given by the test leader predicted a high probability of success. Harvey et al. 2016b demonstrated that a more relaxed lead during "robin", "food", and "pigeons" subtests, and playing in the first replicate of the "tea towel" subtest predicted a high probability of success. In this study, the univariate logistic regression model using scores derived from behaviour on a leash during the "walking" and "distractions" revealed a statistically significant ($p < 0.05$) association with working outcome. Furthermore, both scores were selected for the optimal multivariate logistic regression model; in particular, "S Distractions Pull_mean" in DC1 and "S Walking Pull_mean" in DC2. These variables achieved the significance threshold and were considered the second and first most important features in those multivariate models, respectively. Interestingly, dogs who constantly and strongly pulled on leashes during both subtest 2 "walking" and subtest 8 "distractions" had a lower probability of failing.

Subtests 3 to 5 were similar to the "obedience training" in Weiss et al. 1997 and Weiss 2002, to the "novel situation" test in Weiss 2002, and to

the "obedience" subtest in Harvey et al. 2016b. As mentioned, the main goal of the subtests "standing", "sitting", and "lying down" was to collect inertial data for the posture recognition algorithm developed in Chapter 5. Weiss et al. 1997 and Weiss 2002 did not evaluate the link between behavioural measures related to obedience and training outcomes. Harvey et al. 2016b found a significant association between behaviours shown during the obedience test with the puppy walker at 5 months old only. In particular, there was a positive association between high scores of lip licks and withdrawal and a negative relationship between gaze proportion and withdrawal. In this study, no variables from this subtest showed statistical association with training outcomes and were selected for the predictive model.

Subtest 6 "body check" replicated the "body check" in Harvey et al. 2016b and "physical exam" in Bray et al. 2019, but it was also similar to "restrain" in Asher et al. 2013, as the dog was also gently restrained by the test assistant. The reported results from the analysis of the relationship between the body check subtest with training outcome were mixed. Behavioural measures associated with "restrain" and "body check" were not significant, while body tension and ease of handling in the "physical exam" were significantly associated with training outcomes in assistance dogs. Dogs who were rated lower on body tension (i.e., more relaxed) during the physical tension and higher on ease of handling (i.e., more resistant) had a lower probability of withdrawal. In this research, the composite score associated with "S Sensitivity" was selected for the optimal multivariate model in DC1 and DC2. Similar to Bray et al. 2019, dogs who were willing to jump on the table and comfortable with physical touch had a lower probability of failing in both models.

Subtest 7 "tea towel" was reproduced from Harvey et al. 2016b. At 5 months, dogs were predicted as more likely to be withdrawn if they attempted to remove the tea towel from their back (on each replicate), and played with the tea towel in the second replicate. At 8 months, dogs were predicted as more likely to be withdrawn if they played with the tea towel, changed from the neutral posture in response to the first tea towel replicate, and did not turn to look at the tea towel on their back in the second replicate. In this study, only one ethogram item in DC2 showed statistical significance ($p < 0.05$); specifically, dogs that were indifferent to the tea towel in the second replicate were less likely to fail. No variables from the tea towel were selected for the optimal model.

Similarly, subtest 8 "distractions" was also reproduced from Harvey et al. 2016b. In that study, the only ethogram item significantly associated with training outcome at both 5 and 8 months was the time oriented to the "food". Three additional variables were significantly associated at 8 months only, specifically, they measured whether the dog avoided the "pigeon", pulled towards the "food" and/or "robin". In our

study, the first response to the car (similar to the "robin" subtest) in DC2 was positively associated with working outcomes at traditional levels ($p < 0.05$). The final multivariate model included "S Distractions Pull_mean" in DC1 and "Distractions First Response Car" in DC2, which decreased and increased the probability of withdrawal, respectively.

Subtest 9 "Kong" was similar to the subtests in MacLean et al. 2018 L. Batt et al. 2008a, L. M. Tomkins et al. 2012, and MacLean et al. 2018. Dogs who used both paws more frequently (L. Batt et al. 2008a) or preferred using the right paw L. M. Tomkins et al. 2012 were more successful. MacLean et al. 2018 did not report specific results on the relationship between lateral measures and training outcomes. In our study, no significant statistical difference was found between the two outcomes at ($p < 0.05$) and no variables were selected for the prediction models.

Unlike L. Batt et al. 2008a, L. M. Tomkins et al. 2011, and Bray et al. 2019, this study did not find a significant relationship between the variables recorded during the subtest 10 "dog distraction" and working outcomes. The ethogram items related to this subtest were different from previous studies as they focused on the subject dog's interaction with the distraction dog, handler, and environment, rather than their specific behavioural response.

No ethogram items or composite scores derived from subtest 11 "crate" showed statistical association with working dog outcomes at traditional levels in our study. Nonetheless, the final model included "S Crate Stimulus_mean" in DC1 and "S Crate Handler_mean" in DC2 which increased and decreased the probability of withdrawal, respectively. Hence, dogs that frequently engaged with the crate were more likely to fail and dogs that sought handler attention were less likely to fail. This result is in line with L. M. Tomkins et al. 2011 and contrary to L. Batt et al. 2008a who discovered a significant relationship between training outcome and the latency to drop and rest in the "passive test" with 6-month-old prospective guide dogs.

Subtest 12 "petting" was adapted from Asher et al. 2013, who reported that dogs who were rated as less confident on the stroking stimulus, were more likely to be withdrawn from the guide dog training programme. Similarly, the optimal model in DC1 used "S Petting Stimulus During_mean", which indicated that less playful interactions between handler and dog predicted a higher probability of withdrawal in DC1.

Subtest 13 "isolation" was performed in other studies; however, Bray et al. 2017a did not find a link between ethogram measures and release from the programme, and Wilsson et al. 1998 did not investigate the relationship with training outcome. In this study, a significant relationship was found between two ethogram items and composite

scores measuring time oriented and exploration during this subtest and working outcome in DC2. The optimal model used "S Isolation Stimulus", indicating that dogs who more frequently explored and less frequently paced had a higher probability of withdrawal.

Although similar tests to subtest 14 "noise" were performed in many studies, only L. Batt et al. 2008a, L. M. Tomkins et al. 2012, Asher et al. 2013, and Bray et al. 2019 analysed behavioural measures with training outcome in assistance dogs. Results presented in the literature were controversial, as L. Batt et al. 2008a and Asher et al. 2013 reported no statistical significance, while L. M. Tomkins et al. 2012 found a significant relationship between latency to sit and guide dog success, and Bray et al. 2019 successfully built a model taking several measures derived from the noise subtest to predict training outcome. In our study, only one ethogram item was derived from this subtest, but it did not achieve statistical significance.

Conclusively, it is important to highlight that some subtests present varying degrees of similarity with subtests from previous research. However, there were some important differences which could have significantly impacted results. Even for identical subtests, measuring different ethogram items completely changes the results, especially when testing its relationship with training outcome, as that is also influenced by external factors (L. Batt et al. 2010). Different ethogram measures could be developed for the ADTB in the future, considering the results from this preliminary study to improve prediction results.

#### 4.4.4.2 Prediction Performance

Table 4.14 shows previous studies in the field of training outcome prediction for assistance dogs in order of publication. It outlines the number of dogs, the class imbalance (success rate), age at assessment measured in months, the method used for estimating performance metrics and finally different performance metrics; namely, ROC, precision, sensitivity, specificity, and accuracy. Blanks indicate that the information was not found in the published study.

Table 4.14: Comparison between the performance results achieved by the two classifiers in this study and previously reported research.

| Test Name<br>Publication | N | Imbalance | Age | Method | ROC | Precision | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| This study | | | | | | | | | |
| ADTB (DC1) | 98 | 0.79 | 16 | LOOCV | 0.74 | 0.50 | 0.24 | 0.94 | 0.79 |
| ADTB (DC2) | 60 | 0.80 | 18 | LOOCV | 0.84 | 0.67 | 0.67 | 0.92 | 0.87 |
| Previous studies | | | | | | | | | |
| NA<br>L. Batt et al. 2008a | 43 | 0.53 | 6 | | | | | | 0.27 |
| | 43 | 0.53 | 14 | | | | | | 1.00 |
| | 43 | 0.53 | 14-20 | | | | | | 0.32 |
| PPA<br>Asher et al. 2013 | 465 | 0.64 | 2 | | | 0.93 | 0.08 | 1.00 | |
| JGDBT<br>Harvey et al. 2016b | 73 | 0.71 | 5 | | | | | | 0.80 |
| | 72 | 0.79 | 8 | | | | | | 0.87 |
| YAT<br>Bray et al. 2017a | 98 | 0.68 | 14-17 | | 0.75 | | | | |
| DCTB<br>MacLean et al. 2018 | 111 | 0.68 | 24 | 4-fold | 0.76 | | | | 0.72 |
| | 121 | 0.77 | 24 | Hold out | 0.68 | | | | 0.61 |
| IFT<br>Bray et al. 2019 | 5,967 | 0.42 | 18 | Hold out | 0.71 | | | | 0.66 |

Test Name: ADTB = Assistance Dog Test Battery; NA = Not Available; PPA = Puppy Profiling Assessment; JGDBT = Juvenile Guide Dog Behavior Test; YAT = Young Adult Test; DCTB = Dog Cognition Test Battery; IFT = In For Training. Imbalance: the proportion of examples from the majority class (success). Age: given in months. Method: Leave One Out Cross Validation (LOOCV).

In order to compare the performance metrics of models built in different studies, two characteristics must be considered: the predicted class and the effect of class imbalance on different performance metrics. The predicted class can be success or failure, swapping them causes the precision (also known Positive Predictive Value) to also swap with negative predictive value, similarly sensitivity swaps with specificity. On the other hand, accuracy and ROC remain unchanged. Luque et al. 2019 investigated the influence of class imbalance on different performance metrics. They concluded that sensitivity, specificity, and their geometric mean (g-means) are not affected by imbalance, accuracy is moderately affected by it, and precision and f1-scores are highly biased. The area under the ROC curve is calculated based on the balance between sensitivity and specificity; consequently, it is also considered a reliable metric to compare models built with dissimilar class imbalances. Hence, the area under the ROC curve was used to compare the performance of different models.

The DC1 model developed in this research achieved a ROC of 0.74 which is comparable to the state-of-the-art, while the DC2 attained a ROC of 0.84 outperforming all previously reported models. L. Batt et al. 2008a was able to predict training outcomes in dogs at 14 months with perfect accuracy. However, this was a small sample, and no cross-validation method was reported for the calculation of such a metric. Asher et al. 2013 analysed the potential of setting a cut-off threshold for the composite score, achieving better precision and specificity at the expense of worse sensitivity. Harvey et al. 2016b achieved similar results compared to this study; however, no form of cross-validation was reported. Bray et al. 2017a reported similar performance metrics to DC1; nonetheless, it is not clear which cross-validation method was used for estimating the results. MacLean et al. 2018 had a larger sample of dogs and reported a significantly lower performance in the hold-out test set than in the cross-validated set, due to the dissimilar class imbalances. Bray et al. 2019 used a significantly larger dataset; however, the performance of their predictive model was considerably worse than this study's model created with the DC2 dataset.

Conclusively, this research demonstrated the relationship between novel ADTB ethogram measures and working outcomes in assistance dogs. Moreover, the performance results achieved by the model created using the DC1 dataset equalled the state-of-the-art, while the DC2 dataset substantially outperformed it. These preliminary results suggest that the novel ADTB performed ten weeks after the start of training can provide more reliable predictions of working outcomes than the models published in the literature. This research contributes to the body of knowledge, by providing the most reliable tool for the prediction of working outcomes in trainee assistance dogs. Hence, model predictions can be used by assistance dog organisations to reduce the cost of

training by withdrawing dogs from training.

# Chapter 5

# Posture Estimation

## 5.1 Introduction

As discussed in Chapter 2, a variety of studies assessed the suitability of dogs for assistance work. Behavioural assessment methods can be classified into ratings, test batteries, observational tests, and canine activity monitoring. The most commonly used methods are ratings which were utilised in Chapter 3 and test batteries which were employed in Chapter 4. The work presented in this chapter concerns canine activity monitoring.

This research aims to achieve the secondary objective of this thesis which is to improve the performance of canine posture recognition systems. Smart canine activity monitors have been developed to accurately and reliably monitor activity and health parameters to provide an indicator of overall welfare (Ladha et al. 2013; S. Jones et al. 2014; Belda et al. 2018). In order to achieve that, they need to correctly quantify activity levels and identify important behaviours.

This technological advancement provides a huge opportunity not only for pet owners but also for working dog organisations including guide, assistance, police, search and rescue, etc. In the case of working dogs, activity monitors capable of estimating posture could assist in the behaviour assessment of trainee dogs (Gerencsér et al. 2013; Mealin et al. 2020; Mealin 2021) and allow the development of computer-canine training systems (Brugarolas et al. 2012; J. Majikes et al. 2017; J. J. Majikes et al. 2017). In particular, activities estimated by a posture recognition algorithm during a behaviour test could be used to automatically create ethograms Gerencsér et al. 2013, to predict a dog's behaviour scores during a test battery (Mealin et al. 2020) or to estimate a dog's probability of success in their training programme (Mealin 2021).

Furthermore, posture recognition systems could also be used to measure the level of rest at night which is indicative of training

outcomes in apprentice guide dogs (L. M. Tomkins et al. 2011). The rate of occurrence of static and inactive behaviour instead of sleep was associated with anxiety and hyper-vigilance (den Uijl et al. 2017). Another possible application of posture recognition systems regards the monitoring of welfare as some activity and health parameters have been associated with the occurrence of emotional states and stress in dogs (Bonne Beerda et al. 1997; S. Jones et al. 2014). They could also be utilised for noise cancellation in the interpretation of physiological measurements (Bozkurt et al. 2014; Part et al. 2014).

The main goal of this chapter is to improve the performance of canine posture recognition systems for working dogs. In order to achieve that goal, this study explores the use of multiple inertial sensors, innovative data pre-processing techniques and novel machine learning model architectures. Accordingly, the following research questions were investigated:

**Q.7** Which Inertial Measurement Unit (IMU) sensor position yields the most predictive features for canine posture recognition?

**Q.8** Which feature domain produces the most predictive features for canine posture recognition?

## 5.2 Methodology

### 5.2.1 Subjects

The subjects were 42 healthy apprentice dogs participating in the assistance dog programme at the Irish Guide Dogs for the Blind (IGDB) Training Centre in Cork, Ireland. The dogs were Golden Retrievers (N = 1, Mean Age = 14.6 months), Labrador Retrievers (N = 16, Mean Age = 17.3 month) and Crosses (N = 25, Mean Age = 11.6 months).

### 5.2.2 Devices and Data

Inertial data were gathered during the data collection session by three IMUs GT9X Links (Actigraph, Pensacola, USA) containing a three-axial accelerometer, gyroscope, and magnetometer. ActiLife software (Actigraph, Pensacola, USA) was used to initialise the devices with a 100Hz sampling rate.

The IMUs were attached to fabric straps on the dogs through hook-and-loop fasteners glued to the devices and on the fabric straps (Xsens, Enschede, the Netherlands). The sensors were placed on the dogs' neck, back, and chest following the placement shown in Fig 5.1.

The IMU Raw datasets were concatenated producing 27 data streams, as follows: 3 IMUs (neck, chest, back) containing 3 sensors each

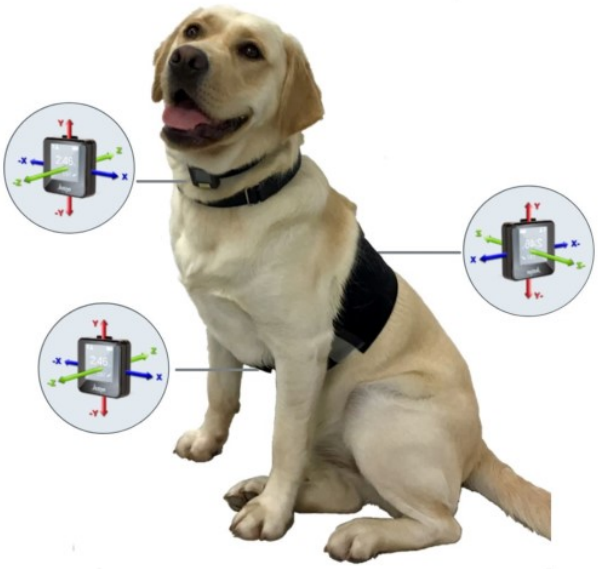


Figure 5.1: IMUs sensor placement on the dogs' neck, back, and chest.

(accelerometer, gyroscope, magnetometer) with 3 axes each (X, Y, Z). The magnetometer features were removed from the dataset as initial experiments confirmed that they were not as predictive of posture. Therefore, the resulting IMU Raw dataset comprised of accelerometer and gyroscope measurements resulting in a total of 18 columns.

### 5.2.3 Data Collection Protocol

The IMUs were initialised using the local time and placed on the dog to gather inertial data on canine postures. In order to synchronise the IMU data and video recording, the local time was shown at the start of the video recording. The postures walking, standing, sitting, and lying down were performed following commands during their respective subtest in the Assistance Dog Test Battery (ADTB) (described in Table 4.1) while body shakes occurred spontaneously during other subtests. Some subjects were more cooperative than others in following the instructions to perform and hold postures for the predetermined period of time. Hence, positive reinforcement methods were administered by the handler using rewards in the form of verbal praise and food to increase adherence to the protocol.

### 5.2.4 Data Annotation

The postures performed by the dog in the video-recorded session were annotated and classified into two types and five postures, as described in Table 5.1.

Posture Timestamps datasets were created for each video-recorded data

Table 5.1: Categories of behaviour analysed.

| Type | Posture |
|---|---|
| Dynamic | Walking, Body shake. |
| Static | Standing, Sitting, Lying down. |

collection session second by second. They comprised the timestamp for the start and finish video recording time for each posture described.

The IMU Raw dataset was combined with the Posture Timestamps dataset to originate a unified annotated dataset entitled IMU Posture dataset which contains IMU Raw data, dog name and breed, data collection number and the two labels; namely, posture and type. Transitions and miscellaneous body postures which did not fit any of the previous categories were manually excluded from the dataset. Table 5.2 shows the number of observations in the IMU Posture dataset per posture and type.

Table 5.2: Type, posture, and number of observations in the IMU Posture dataset.

| Type | Observations | Posture | Observations |
|---|---|---|---|
| Dynamic | 475,802 | Walking | 466,502 |
| | | Body shake | 9,300 |
| Static | 1,199,200 | Standing | 640,700 |
| | | Sitting | 323,200 |
| | | Lying down | 235,300 |
| Total | | | 1,675,002 |

### 5.2.5 Feature Extraction

Features were calculated using the IMU Posture dataset considering windows of time where a unique posture was performed. Three parameters control the creation of the window as illustrated in Fig 5.2, and described below:

- Transition time (t_time): time between different postures: t_time = $0.25s$.
- Window size (w_size): rolling window size for calculating statistical measures: w_size = $1s$.
- Window offset (w_offset): sampling offsets in the rolling window: w_offset = $0.5s$.

The dataset was created by calculating diverse statistical, temporal and spectral measures on rolling windows controlled by the re-sampling

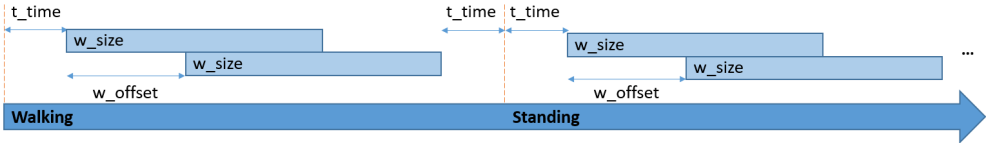



Figure 5.2: Rolling window used for selecting data for feature extraction. Hyper-parameters shown include re-sampling transition time (t_time) between postures, window size (w_size) used to calculate the statistical measures and window offset (w_offset) to create separate observations.

hyper-parameters described above, following the standard 50% overlap between subsequent frames (Ladha et al. 2013). The feature set was obtained with tsfel package (Barandas et al. 2020) which calculated 185 different variables for each of the 18 original raw inertial signals in the IMU Posture dataset, resulting in 3,330 features in the final set. These 185 features calculated by tsfel belonged to 60 types, including 26 spectral, 18 temporal, and 16 statistical features.

### 5.2.6 Classification Models

All classification models were developed using Python (v3.8.1), in particular, tsfel (v0.1.4) was used to extract features and sklearn (v1.0.2) and imblearn (v0.9.0) libraries were used to implement machine learning techniques. The models were trained on ICHEC's Kay supercomputer nodes comprising of 2 x 20 core 2.4GHz Intel Xeon Gold 6148 processors (ICHEC 2022).

The original dataset was split into three sets named development, GR, and test while controlling for the dogs in order to prevent overlap between datasets. Feature extraction was performed with the Python package tsfel (Barandas et al. 2020) using a rolling window, resulting in a dataset composed of 3,300 features and 30,224 observations. The observation counts for each of these datasets after feature extraction using the rolling window are shown in Table 5.3.

Table 5.3: Number of observations after feature extraction per dataset per posture.

| Posture | Development | Test | GR | Total |
|---|---|---|---|---|
| Body shake | 82 | 18 | 2 | 102 |
| Lying down | 3,884 | 384 | 138 | 4,406 |
| Sitting | 5,096 | 946 | 156 | 6,198 |
| Standing | 9,286 | 2,158 | 588 | 12,032 |
| Walking | 6,418 | 1,360 | 398 | 8,176 |
| Total | 24,766 | 4,866 | 1,282 | 30,914 |

The development set contained 36 dogs and comprised 82% of the observations. The test set contained 5 dogs and consisted of 16% of the

observations and was created prioritising dogs who performed all the postures. The GR dataset contained data from the Golden Retriever dog only and represented the remaining 2% of the observations. The only dataset used for training models was the development set, the GR and test sets were used to analyse the intra-breed and inter-breed performance of the different models.

The development set was partitioned to create independent training and validation sets using 10-fold group cross-validation to ensure that those sets did not contain data from the same dogs. The resulting partitioned datasets were used to fit classification models in order to optimise the estimator hyper-parameters.

Select K Best method was utilised to select the best features by calculating ANOVA F-value between the features and labels. Scores given were grouped by sensor position and type to estimate their importance. Random Forest classifier was deployed to fit the development set using all features selected by Select K Best. Feature importance was grouped by sensor position and type, and also feature domain and type.

A grid search algorithm was utilised to evaluate all combinations of the hyper-parameters shown in Table 5.4. In particular, the feature selection algorithm was set to use the top K = 10, 20, 35, 55, and 80 features out of a total of 3,300 initial features extracted by tsfel. It also optimised two hyper-parameters of the Random Forest classifier; namely, maximum depth = 3, 5, 7, 10 and number of estimators = 25, 50, 100, 250, 500.

Table 5.4: Grid search hyper-parameter set for the classifiers.

| **Method** | **Hyper-parameters** |
|---|---|
| Select K Best | k = [10, 20, 35, 55, 80] |
| Random Forest | max_depth = [3, 5, 7, 10]<br>n_estimators = [25, 50, 100, 250, 500] |

The metric chosen to optimise model performance was f1-weighted because it provides a good estimate of the model's ability to predict the correct class in multi-class classification problems with a natural class imbalance. Hence, the grid search chose the hyper-parameter set that yielded the best f1-weighed score. The model using the optimal hyper-parameter set was fit to the entire development set and then used for predicting the label on unseen data from the test set. Intra-breed and inter-breed classification performances were evaluated using the test set (Labrador Retriever and Crosses) and the GR set (Golden Retriever only), respectively.

Besides the optimisation metric f1-weighted, other additional metrics were reported to facilitate the comparison between classifiers presented in this study and classifiers developed in other studies on canine

posture recognition. In particular, the following metrics were reported for each of the postures: Sensitivity also known as True Positive Rate (TPR) or Recall; Specificity also known as True Negative Rate (TNR); Accuracy; Precision also known as Positive Predictive Value (PPV); and f1 score. The following metrics were reported for each of the classifiers: f1-weighted, f1-macro and confusion matrix. Confusion matrices were normalised considering the number of true instances in each class, i.e., the total count in the rows of the matrix.

The next subsections outline different model architectures used for predicting 5 postures; namely, body shake, lying down, sitting, standing, and walking. Classifier 1 comprised one estimator to predict the postures directly. Classifier 2 was a cascade classifier composed of 3 estimators, where the first one detected the type of posture (dynamic or static) and the other two further classified the posture. Classifier 3 had two estimators in sequence, where the first one detected body shakes as an anomaly and the second one classified the remaining 4 postures.

#### 5.2.6.1 Classifier 1

Random Forest classifier was fitted to the entire development dataset with all the 5 classes in order to predict posture directly as indicated in Fig 5.3.

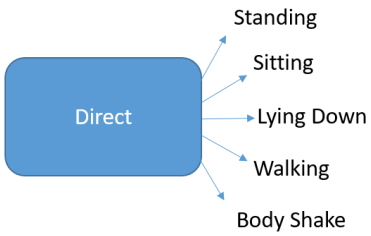



Figure 5.3: Direct estimator used in Classifier 1.

#### 5.2.6.2 Classifier 2

This classifier comprised three estimators, namely, type, static and dynamic, as shown in Fig 5.4. Each estimator was trained separately, as outlined below:

- **Type**: Random Forest classifier was fitted to the entire development dataset to predict the label 'type' as described in Table 5.1 which comprised of two classes; namely, static and dynamic.
- **Static**: Random Forest classifier was fitted to a subset of the dataset containing static postures to predict the classes lying down, sitting and standing.

- **Dynamic**: Random Forest classifier was fitted to a subset of the dataset containing dynamic postures to predict the classes walking and body shake.

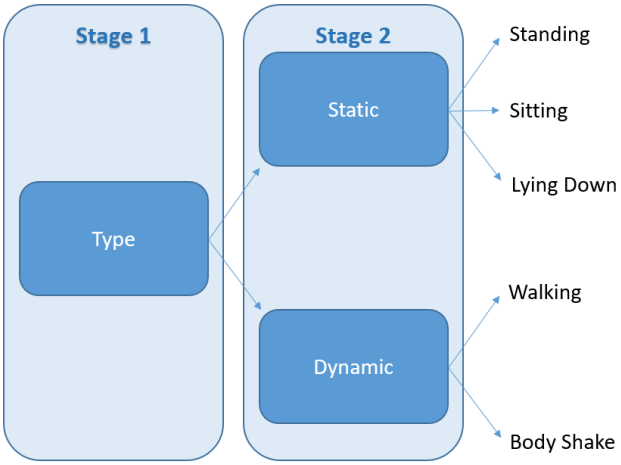



Figure 5.4: Type, Static, and Dynamic estimators used in Classifier 2.

At prediction time, they were arranged sequentially in two stages to estimate the final posture as illustrated in the diagram in Fig 5.4. The class predicted by the type model in the first stage determined which estimator would be used in the second stage.

#### 5.2.6.3 Classifier 3

This classifier comprised two estimators, namely, Anomaly and Normal as indicated in Fig 5.5. In the first stage, it identified abnormal behaviour (body shake). Other instances went to the second stage where the Normal estimator classified all the other four postures in the second stage.

- **Anomaly**: Isolation Forest, instead of Random Forest, was fitted to the entire dataset to identify the minority class body shake while all other four postures were assigned to the majority class. The hyper-parameter space searched for comprised of the following: number of estimators = 25, 50, 100, 250, 500 and contamination = 0.005, 0.01, 0.05, 0.1.
- **Normal**: Random Forest classifier was fitted to a subset of the dataset containing static postures to predict the classes lying down, sitting, standing and walking.

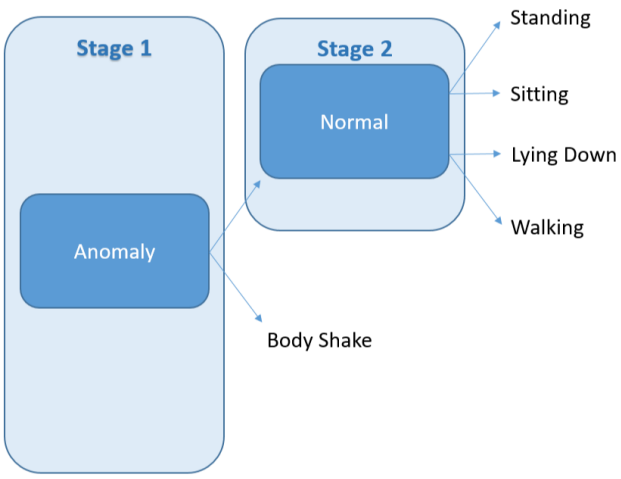



Figure 5.5: Anomaly and Normal estimators used in Classifier 3.

## 5.3 Results

Section 5.3.1 shows the grid search's chosen best model hyper-parameter combination, cross-validation results on the development set and performance on the test set evaluated individually and independently. Feature contribution was evaluated based on Select K Best's f-classification score and Random Forest classifier's feature importance metric in Classifier 1. Section 5.3.2 presents the grouped figures as an estimate of the collective contribution of each sensor type and position.

The best models were deployed as indicated in the classifier diagram in Fig 5.3, 5.4, and 5.5. They were evaluated on the test set to estimate intra-breed performance in Section 5.3.3 and the GR set to estimate inter-breed performance in Section 5.3.4. The performance achieved using the best classifier to predict posture on the combined test and GR sets is shown in Section 5.3.5.

### 5.3.1 Classification Models

#### 5.3.1.1 Classifier 1

The best hyper-parameters chosen for the Direct estimator were K = 80 for Select K Best and maximum depth = 10 and number of estimators = 250 for Random Forest. This model achieved an f1-weighted mean of 0.89 $\pm 0.04$ on the validation set, and an f1-weighted 0.89 and f1-macro 0.82 on the test set.

#### 5.3.1.2 Classifier 2

The best hyper-parameters chosen for the Type, Dynamic, and Static estimators are shown in Table 5.5 along with the f1-weighted

performance metric for the validation and test sets.

Table 5.5: Classifier 2 model's hyper-parameter set selected by grid search and f1-weighted performance on 10-fold validation sets (mean ± standard deviation) and test set.

| Estimator | Hyper-parameters | Validation | Test |
|---|---|---|---|
| Type | k = 80<br>max_depth = 10<br>n_estimators = 100 | $0.95 \pm 0.02$ | 0.95 |
| Dynamic | k = 10<br>max_depth = 10<br>n_estimators = 25 | $0.98 \pm 0.01$ | 0.99 |
| Static | k = 80<br>max_depth = 10<br>n_estimators = 500 | $0.93 \pm 0.03$ | 0.93 |

#### 5.3.1.3 Classifier 3

The best hyper-parameters chosen by grid search for the Anomaly and Normal estimators are shown in Table 5.6 along with the f1-weighted performance metric for the validation and test set.

Table 5.6: Classifier 3 model's hyper-parameter set selected by grid search and f1-weighted performance on 10-fold validation sets (mean ± standard deviation) and test set.

| Estimator | Hyper-parameters | Validation | Test |
|---|---|---|---|
| Anomaly | k = 10<br>n_estimators = 50<br>contamination = 0.005 | $0.99 \pm 0.03$ | 0.95 |
| Normal | k = 80<br>max_depth = 10<br>n_estimators = 500 | $0.90 \pm 0.03$ | 0.90 |

### 5.3.2 Feature Importance

Select K Best used f-classification to calculate the f-score and p-value for each feature considering the posture label. The score given was then normalised and grouped by sensor position (back, chest, and neck) and sensor type (accelerometer and gyroscope) to estimate their collective contribution to the model. These results are shown in Table 5.7 and indicate that the most important sensor position is the back and the most important sensor type is the accelerometer for distinguishing between different postures.

Table 5.7: Feature scores calculated by f-classification in Select K Best by IMU Position and Sensor.

| Position | SKB Score | Type | SKB Score |
|---|---|---|---|
| Back | 0.45 | Acc | 0.26 |
| | | Gyr | 0.19 |
| Chest | 0.39 | Acc | 0.22 |
| | | Gyr | 0.17 |
| Neck | 0.16 | Acc | 0.09 |
| | | Gyr | 0.07 |

Table 5.8: Feature importance calculated by Random Forest classifier considering the 80 features previously selected by Select K Best.

| Position | RF Importance | Type | RF Importance |
|---|---|---|---|
| Back | 0.57 | Acc | 0.37 |
| | | Gyr | 0.20 |
| Chest | 0.43 | Acc | 0.24 |
| | | Gyr | 0.19 |

Table 5.8 shows the feature importance given by the Random Forest from Classifier 1 on the 80 features chosen by Select K Best. No features derived from the neck sensor were among the 80 highest-ranked ones. Once again, these results indicate that the most important sensor position was back and the sensor type was the accelerometer.

Table 5.9 shows more information on the 80 highest-scoring features chosen by Select K best including their domain and collective importance. They belonged to 26 feature types out of the 60 domains calculated, specifically, there were 13 out of 26 spectral, 7 out of 18 temporal, and 5 out of 16 statistical features. In order of importance, the domains providing the best features were the statistical, temporal and spectral domains. Even though the spectral domain had the biggest number of statistical feature types selected, it provided the least collective importance as calculated by Random Forest.

### 5.3.3 Intra-Breed Evaluation

This section refers to the performance evaluation of the model when predicting the posture of dogs belonging to the same breed as that used in the creation of the model (i.e., intra-breed). The test set was used to evaluate the performance of each experimental model trained on data from Labrador Retrievers (pure and Golden Retriever crosses) on dogs from the same breed. A summary of the classification metrics on the test set for Classifiers 1, 2, and 3 per class and f1-weighted and f1-macro averages is shown in Tables 5.10, 5.11 and 5.12, respectively. The three classifiers achieved very similar results. Classifier 3 achieved the best

Table 5.9: Feature importance calculated by Random Forest classifier considering the 80 features selected by Select K Best by domain.

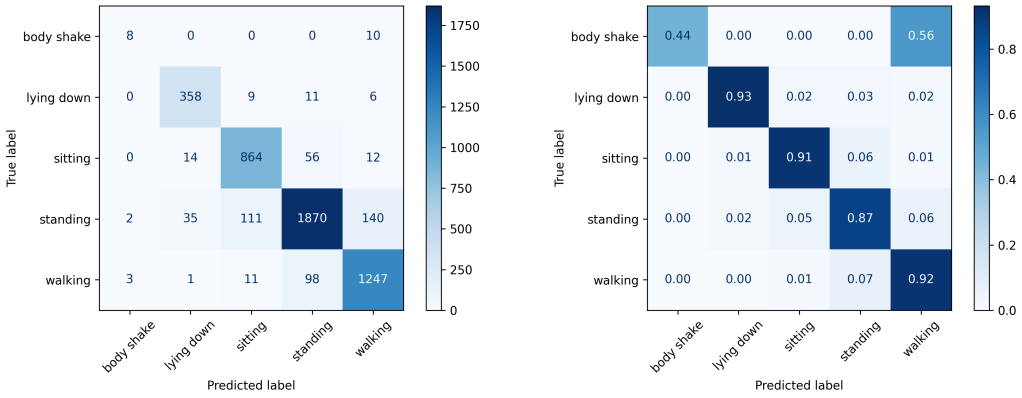
| Domain (N) | Importance | Feature Type |
|---|---|---|
| Statistical (8) | 0.47 | Root mean square, median absolute deviation, empirical cumulative distribution function percentile, interquartile range, minimum, standard deviation, histogram, mean absolute deviation |
| Temporal (5) | 0.32 | Area under the curve, peak to peak distance, negative turning points, entropy, positive turning points |
| Spectral (13) | 0.21 | wavelet absolute mean, wavelet standard deviation, spectral distance, wavelet energy, linear prediction cepstral coefficients, spectral skewness, spectral decrease, spectral centroid, spectral slope, spectral kurtosis, maximum frequency, wavelet variance, spectral roll-off |

f1-macro as a result of its improved ability to correctly identify body shakes; however, its f1-weighted was slightly lower for standing and lying down.

The normalised confusion matrices with the predicted labels on the test set using Classifier 1, 2, and 3 models are shown in Fig 5.6, 5.7, and 5.8, respectively. The main difference appears in the body shake performance as this class was mainly correctly classified by Classifier 3 model only. This improved performance on the minority class did not negatively affect the model's ability to identify other classes as they showed comparable figures to Classifiers 1 and 2.

Table 5.10: Classification metrics per posture achieved using the best models selected by grid search in Classifier 1 on the test set.

| Posture | Sensitivity | Specificity | Accuracy | Precision | F1-score |
|---|---|---|---|---|---|
| Body shake | 0.44 | 1.00 | 1.00 | 0.62 | 0.52 |
| Lying down | 0.93 | 0.99 | 0.98 | 0.88 | 0.90 |
| Sitting | 0.91 | 0.97 | 0.96 | 0.87 | 0.89 |
| Standing | 0.87 | 0.94 | 0.91 | 0.92 | 0.89 |
| Walking | 0.92 | 0.95 | 0.94 | 0.88 | 0.90 |
| f1-weighted | | | 0.89 | | |
| f1-macro | | | 0.82 | | |

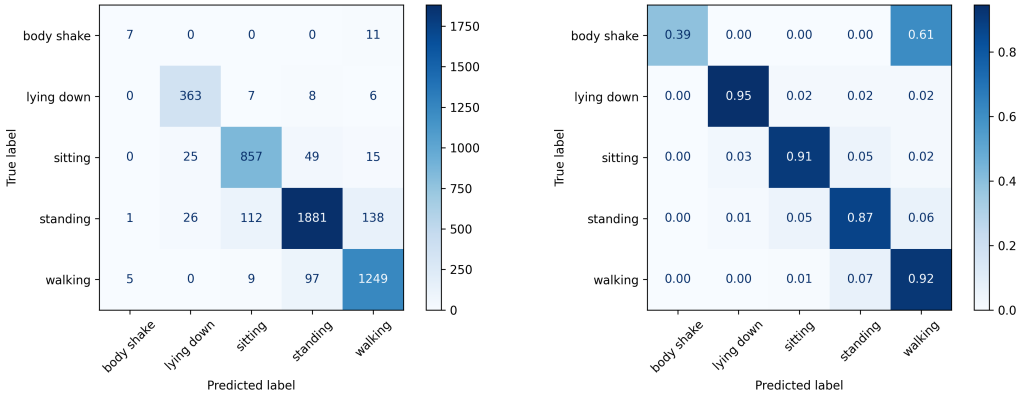



(a) Absolute numbers. (b) Normalised by true label.

Figure 5.6: Confusion matrix for Classifier 1 on the test set.

Table 5.11: Classification metrics per posture achieved using the best models selected by grid search in Classifier 2 on the test set.

| Posture | Sensitivity | Specificity | Accuracy | Precision | F1-score |
|---|---|---|---|---|---|
| Body shake | 0.39 | 1.00 | 1.00 | 0.54 | 0.45 |
| Lying down | 0.95 | 0.99 | 0.99 | 0.88 | 0.91 |
| Sitting | 0.91 | 0.97 | 0.96 | 0.87 | 0.89 |
| Standing | 0.87 | 0.94 | 0.91 | 0.92 | 0.90 |
| Walking | 0.92 | 0.95 | 0.94 | 0.88 | 0.90 |
| f1-weighted | | | 0.90 | | |
| f1-macro | | | 0.81 | | |




(a) Absolute numbers. (b) Normalised by true label.

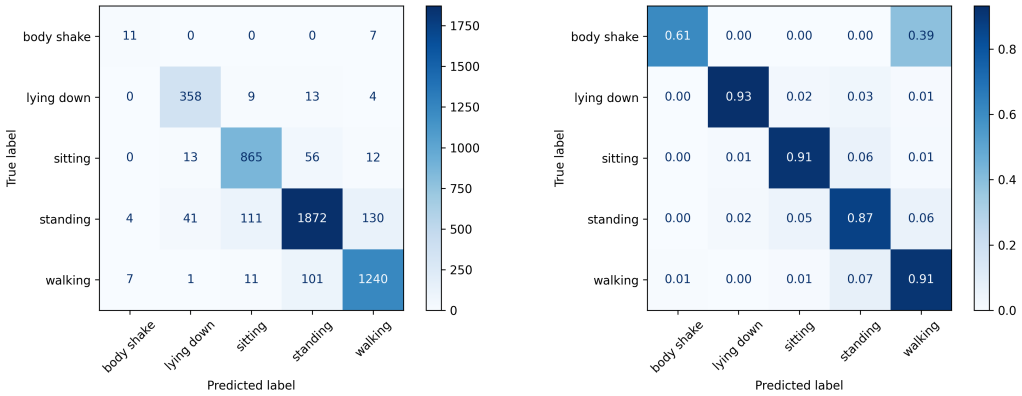
Figure 5.7: Confusion matrix for Classifier 2 on the test set.

### 5.3.4 Inter-Breed Evaluation

This section presents the performance evaluation of the model when predicting the posture of dogs belonging to different breeds to that used in the creation of the model (i.e., inter-breed). The GR dataset was used

Table 5.12: Classification metrics per posture achieved using the best models selected by grid search in Classifier 3 on the test set.

| Posture | Sensitivity | Specificity | Accuracy | Precision | F1-score |
|---|---|---|---|---|---|
| Body shake | 0.61 | 1.00 | 1.00 | 0.50 | 0.55 |
| Lying down | 0.93 | 0.99 | 0.98 | 0.87 | 0.90 |
| Sitting | 0.91 | 0.97 | 0.96 | 0.87 | 0.89 |
| Standing | 0.87 | 0.94 | 0.91 | 0.92 | 0.89 |
| Walking | 0.91 | 0.96 | 0.94 | 0.89 | 0.90 |
| f1-weighted | | | 0.89 | | |
| f1-macro | | | 0.83 | | |




(a) Absolute numbers.

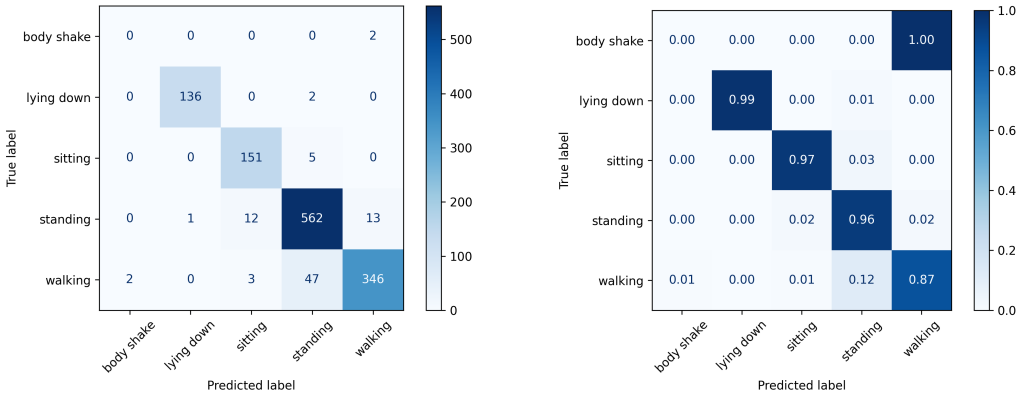
(b) Normalised by true label.

Figure 5.8: Confusion matrix for Classifier 3 on the test set.

to evaluate the performance of each experimental model trained on data from Labrador Retrievers (pure and Golden Retriever crosses) on dogs from another breed (Golden Retriever).

A summary of the classification metrics on the GR set for Classifiers 1, 2, and 3 per class, and f1-weighted and f1-macro averages are shown in Table 5.13, 5.14, and 5.15, respectively. Classifier 3 again achieved the best f1-macro and f1-weighted metrics, indicating that it was the only model capable of correctly classifying body shakes.

The absolute and normalised confusion matrices with the predicted labels for the GR set using Classifier 1, 2, and 3 models are shown in Fig 5.9, 5.10, and 5.11, respectively. It can be visually seen that Classifier 3 achieves the overall best results in correctly classifying the classes. It is important to consider that these results were obtained using the GR dataset, which contained one subject.

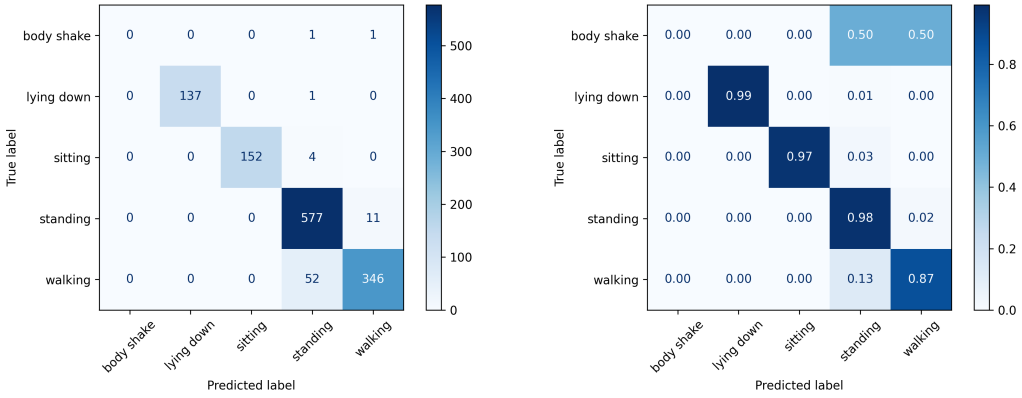



(a) Absolute numbers. (b) Normalised by true label.

Figure 5.9: Confusion matrix for Classifier 1 on GR set.

Table 5.13: Classification metrics per posture achieved using the best models selected by grid search in Classifier 1 on GR set.

| Posture | Sensitivity | Specificity | Accuracy | Precision | F1-score |
|---|---|---|---|---|---|
| Body shake | 0.00 | 1.00 | 1.00 | 0.00 | 0.00 |
| Lying down | 0.99 | 1.00 | 1.00 | 0.99 | 0.99 |
| Sitting | 0.97 | 0.99 | 0.98 | 0.91 | 0.94 |
| Standing | 0.96 | 0.92 | 0.94 | 0.91 | 0.93 |
| Walking | 0.87 | 0.98 | 0.95 | 0.96 | 0.91 |
| f1-weighted | | | 0.93 | | |
| f1-macro | | | 0.75 | | |




(a) Absolute numbers. (b) Normalised by true label.

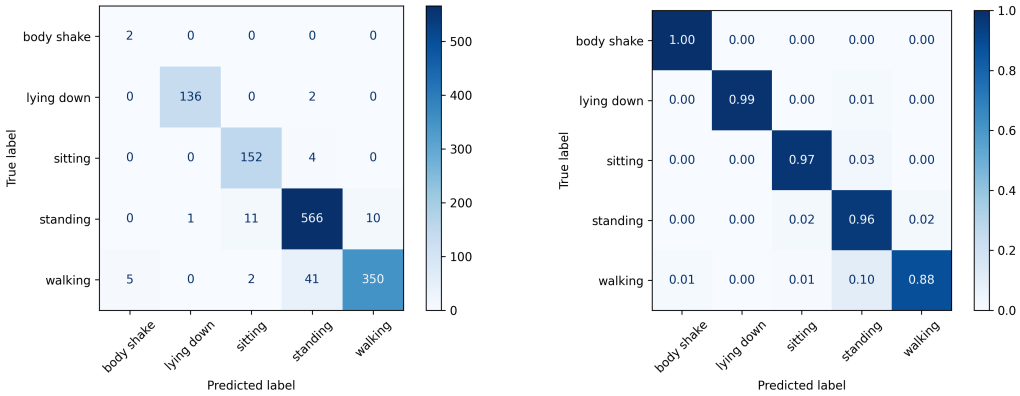
Figure 5.10: Confusion matrix for Classifier 2 on GR set.

### 5.3.5 Best Classifier

Intra-breed and inter-breed evaluation revealed that there were no significant performance differences between Golden Retrievers, Labrador Retrievers and their crosses. Classifier 3 was selected as the

Table 5.14: Classification metrics per posture achieved using the best models selected by grid search in Classifier 2 on GR set.

| Posture | Sensitivity | Specificity | Accuracy | Precision | F1-score |
|---|---|---|---|---|---|
| Body shake | 0.00 | 1.00 | 1.00 | 0.00 | 0.00 |
| Lying down | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |
| Sitting | 0.97 | 1.00 | 1.00 | 1.00 | 0.99 |
| Standing | 0.98 | 0.92 | 0.95 | 0.91 | 0.94 |
| Walking | 0.87 | 0.99 | 0.95 | 0.97 | 0.92 |
| f1-weighted | | | 0.94 | | |
| f1-macro | | | 0.76 | | |




(a) Absolute numbers.

(b) Normalised by true label.

Figure 5.11: Confusion matrix for Classifier 3 on GR set.

Table 5.15: Classification metrics per posture achieved using the best models selected by grid search in Classifier 3 on GR set.

| Posture | Sensitivity | Specificity | Accuracy | Precision | F1-score |
|---|---|---|---|---|---|
| Body shake | 1.00 | 1.00 | 1.00 | 0.29 | 0.44 |
| Lying down | 0.99 | 1.00 | 1.00 | 0.99 | 0.99 |
| Sitting | 0.97 | 0.99 | 0.99 | 0.92 | 0.95 |
| Standing | 0.96 | 0.93 | 0.95 | 0.92 | 0.94 |
| Walking | 0.88 | 0.99 | 0.95 | 0.97 | 0.92 |
| f1-weighted | | | 0.94 | | |
| f1-macro | | | 0.85 | | |

best classifier considering the results from Section 5.3.1. In order to obtain unified performance metrics for the best model, Classifier 3 was used to calculate performance on the combined test and GR sets.

A summary of the key classification metrics per class and f1-weighted and f1-macro averages is shown in Table 5.16, and a confusion matrix is shown in Fig 5.12. These are considered the final results of the present work.

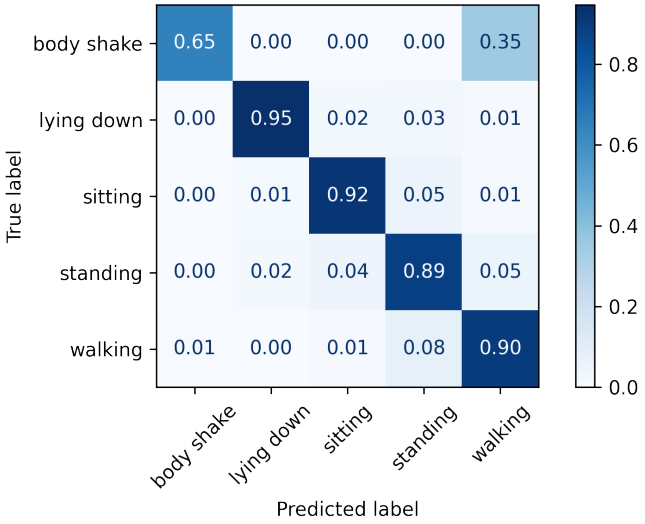



Figure 5.12: Confusion matrix for Classifier 3 on test and GR sets.

The minority class body shake was sometimes incorrectly classified as walking. Such results can be attributed to two main causes. Firstly, it is reasonable to estimate that a significant number of miss-classification cases could have resulted from inaccurate labels. This is because body shakes are very rapid movements typically happening between walking or standing postures. As annotations were made second by second, some frames may contain mixed postures. Secondly, these are the only dynamic postures while all others are static. It can be assumed that there exists a higher level of similarity between these postures when compared to all other postures, making it harder to distinguish them.

Table 5.16: Classification metrics per posture achieved using the best models selected by grid search in Classifier 3 on test and GR sets combined.

| Posture | Sensitivity | Specificity | Accuracy | Precision | F1-score |
|---|---|---|---|---|---|
| Body shake | 0.65 | 1.00 | 1.00 | 0.45 | 0.53 |
| Lying down | 0.95 | 0.99 | 0.99 | 0.90 | 0.92 |
| Sitting | 0.92 | 0.97 | 0.96 | 0.88 | 0.90 |
| Standing | 0.89 | 0.94 | 0.91 | 0.92 | 0.90 |
| Walking | 0.90 | 0.96 | 0.95 | 0.91 | 0.91 |
| f1-weighted | | | 0.90 | | |
| f1-macro | | | 0.83 | | |

## 5.4 Discussion

The present work advanced the state-of-the-art by developing a novel canine posture estimation system capable of classifying five postures

(body shake, lying down, sitting, standing, and walking). The new system was specifically designed for commonly used working dog breeds with the purpose of improving the rate of correct canine posture classifications. Accordingly, it comprised three IMUs placed on the neck, back, and chest to leverage the fact these canines wear a harness while working outdoors. Advanced feature extraction for time series methods and feature selection techniques were applied for the first time to a canine posture recognition system based on inertial sensing. Moreover, three different model architectures were experimented with and grid search was utilised to optimise the models.

This work contributes to the body of knowledge in using IMUs for posture prediction using machine learning by expanding the current understanding of sensor position, feature extraction and importance as well as model architecture. Advanced feature extraction techniques for time series were successfully employed and validated for the first time in this canine posture prediction application. Feature selection was necessary to remove uninformative features, reducing the complexity of models and processing time.

The analysis of the collective contributions of the features deriving from each of the three IMUs placed on the dog's neck, chest, and back revealed the importance of the back and chest sensors for the first time. Features were chosen by the Select K Best algorithm based on the scores given by f-classification considering the label posture. Only features from the back and chest IMUs were selected by the algorithm for the final posture estimation model; consequently, the neck sensor measures were completely removed from the present model. Hence, these results indicate the importance of adding IMU sensors to dog harnesses for applications, where the accuracy of posture predictions is critical, as in the case of some types of working dogs.

Accelerometer data were more indicative of posture than gyroscope data as shown in Tables 5.7 and 5.8 which is in line with previous research (Gerencsér et al. 2013; Brugarolas et al. 2013a). It is important to note that these results are influenced by the types of posture targeted in each study.

A grid search was deployed to find the optimal combination of hyper-parameters for the feature selection and classifier. In particular, it used a 10-fold group cross-validation technique to calculate the performance in 10 different subsets of the development set while controlling for dogs to select the hyper-parameters that resulted in the best f1-weighted performance.

Three classifier architectures were evaluated in order to improve the overall performance in correctly identifying postures despite the naturally unequal distribution of observations per label. Modifying the classifier architecture resulted in an improvement of the f1-score of the

minority class (body shake) while maintaining a comparable performance on the majority classes (lying down, sitting, standing, and walking). This can be particularly observed from Classifier 1 and 3 model results in Tables 5.10 and 5.12 on the test set and Tables 5.13 and 5.15 on the GR set. In the latter case, Classifiers 1 and 2 did not appropriately identify the body shake at all while Classifier 3 correctly labelled all body shakes. The best-performing model was Classifier 3 which comprised an anomaly detection estimator to identify the minority class (body shake) followed by a classifier to detect the other 4 postures (lying down, sitting, standing, and walking). The noticeably higher performance proves the advantage of combining different classifiers to leverage their respective strengths.

One unexpected result was the superior performance of all classifiers on the GR set as compared to the test set. The only class performing more poorly was body shake as indicated by the f1-scores in Tables 5.10, 5.11 and 5.12 on the test and Tables 5.13, 5.14 and 5.15 on the GR set. Interestingly, Gerencsér et al. 2013 also reported a similar result as the inter-breed model achieved higher performance than the intra-breed, from 70.3% to 73.6%, and 72.6% to 73.5% using only Malinois and Labradors data for training, respectively. This indicates that the variability deriving from the breed difference may be less significant than the methodological limitations of the data collection and annotation procedures. Hence, this result suggests that this model could be successfully utilised on not only Labrador Retrievers and crosses with Golden Retrievers but also pure Golden Retrievers as well. Further experiments on more Golden Retrievers are recommended since the GR dataset contained only one subject.

The canine posture estimator developed in this study achieved the best overall performance for predicting the five postures compared to the literature. This outcome was produced by the use of inertial sensors, advanced feature extraction techniques for time series data, feature selection methods and optimal model architecture. The dataset comprising of raw labelled inertial data from 3 IMUs on 42 dogs is made publicly available. Hence, this research contributes to the body of knowledge by providing not only an effective tool for posture monitoring but also a full dataset which allows the application of other machine learning techniques and performance comparison.

In comparison with previous studies, this study advances the state of the art by (1) providing the largest open access dataset for dog activity recognition specific to predominant assistance dog breeds; namely, Labrador Retrievers, Golden Retrievers and their crosses (Ennik et al. 2006; K. M. Evans et al. 2015; van Rooy et al. 2019); (2) applying advanced time series feature extraction methods for the first time in this field; (3) evaluating three different machine learning architectures to address the natural class imbalance; (4) predicting body shake in

addition to the most common dog postures (walking, sitting, standing and lying down); (5) investigating the effect of IMU placement (neck, back and chest). The annotated dataset and code used in this study are publicly available on Mendeley Data and GitHub, respectively.

### 5.4.1 Comparison with Previous Research

Classifier 3 was chosen as the best classification model in the present work and its performance on the test and GR sets combined as shown in Section 5.3.5 was compared to previous studies in Table 5.17. The inclusion criteria for such comparison considered only papers that (1) estimated canine posture using IMUs; (2) provided classification metrics per posture (body shake, lying down, sitting, standing, and walking); (3) evaluated performance on different dogs due to significant performance difference between intra and inter-dog metrics (Brugarolas et al. 2013a; Gerencsér et al. 2013). Considering the nine research papers in Table 2.3, five papers were excluded because they did not meet the criteria (Ladha et al. 2013; Brugarolas et al. 2013a; Brugarolas et al. 2013b; Kiyohara et al. 2015; Kumpulainen et al. 2018). Papers that used commercial systems were also included (Chambers et al. 2021; den Uijl et al. 2016; den Uijl et al. 2017).

Chambers et al. 2021 recommended the use of sensitivity and specificity to compare the performance of classifiers trained on different datasets in order to control for class imbalance effects. Hence, both metrics were combined in a geometric mean (g-mean) for ease of comparison. The highest g-mean value for each posture was highlighted in bold.

Classifier 3 achieved the highest performance for the postures lying down, sitting and standing on the combined test and GR sets. The best g-means for body shake and walking were reported by Chambers et al. 2021 and den Uijl et al. 2017, respectively. Gerencsér et al. 2013 achieved the highest sensitivity for walking; however, there was insufficient information to calculate other performance metrics. It is important to note that, although Ribeiro et al. 2009 reported a high sensitivity for sitting and standing, there were just a few observations from 5 dogs. Moreover, it was not clear whether the algorithm was built using data from the same dogs that were used in the test set.

Table 5.17: Comparison between Classifier 3 performance and previous studies reporting inter-subject classification performance metrics per posture. The best geometric means between sensitivity and specificity (g-mean) are in bold.

| Publication | Posture | Sensitivity | Specificity | Precision | F1-score | G-mean |
|---|---|---|---|---|---|---|
| This study | Body shake | 0.65 | 1.00 | 0.45 | 0.53 | 0.81 |
| | Lying down | 0.95 | 0.99 | 0.90 | 0.92 | **0.97** |
| | Sitting | 0.92 | 0.97 | 0.88 | 0.90 | **0.95** |
| | Standing | 0.89 | 0.94 | 0.92 | 0.90 | **0.91** |
| | Walking | 0.90 | 0.96 | 0.91 | 0.91 | 0.93 |
| Chambers et al. 2021 | Body shake | 0.916 | 1.000 | 0.795 | 0.851 | **0.957** |
| | Lying down | 0.826 | 0.913 | 0.724 | 0.772 | 0.868 |
| | Sitting | 0.409 | 0.915 | 0.347 | 0.375 | 0.612 |
| | Standing | 0.793 | 0.900 | 0.916 | 0.850 | 0.845 |
| | Walking | 0.903 | 0.969 | 0.706 | 0.792 | 0.935 |
| Kumpulainen et al. 2021 | Lying down | 0.837 | | | | |
| | Sitting | 0.899 | | | | |
| | Standing | 0.941 | | | | |
| | Walking | 0.945 | | | | |
| den Uijl et al. 2017 | Walking | 0.91 | 0.91 | 0.99 | 0.95 | 0.91 |
| den Uijl et al. 2016 | Walking | 0.95 | 0.96 | 0.83 | 0.89 | **0.95** |

Continued on next page.

| Publication | Posture | Sensitivity | Specificity | Precision | F1-score | G-mean |
|---|---|---|---|---|---|---|
| Gerencsér et al. 2013 | Lying down | 0.722 | | | | |
| | Sitting | 0.870 | | | | |
| | Standing | 0.767 | | | | |
| | Walking | 0.975 | | | | |
| Ribeiro et al. 2009 | Lying Down | 0.68 | | | | |
| | Sitting | 0.92 | | | | |
| | Standing | 1.00 | | | | |
| | Walking | 0.36 | | | | |

Classifier 3 achieved the top overall performance on the five postures considering that it attained an arithmetic mean of the f1-score, also known as f1-macro, of 0.83 compared to 0.73 in Chambers et al. 2021. Furthermore, Classifier 3 produced an average sensitivity of 0.92 compared to 0.905, 0.83, 0.74 in Kumpulainen et al. 2021, Gerencsér et al. 2013 and Ribeiro et al. 2009, respectively, taking into account the postures lying down, sitting, standing, and walking. Conclusively, Classifier 3 has outperformed previous models published in the literature for predicting the five postures on dogs that did not belong to the original training set.

### 5.4.2 Limitations and Opportunities

As discussed, inaccurate annotations could account for a significant number of errors. Two main methodological limitations were identified to affect the precision of labels. Firstly, data synchronisation was done by showing the standard time on camera at the beginning of the video recording while this could be improved by shaking the sensor in front of the camera. Secondly, data annotation was done second by second while it could be improved by using some video annotation software to increase the precision of the timestamps. These limitations can result in labels being misaligned with the data.

Other two causes of incorrect labels were identified: undefined transitions between two activities and rapid changes from one activity to another. The former case refers to the variability in the duration of transitions (Ribeiro et al. 2009), for instance, transitioning from walking to standing is nearly instantaneous; however, transitioning from standing to lying down can take a couple of seconds, such an observation could understandably be classified as sitting. The latter is observed in cases where the dog is walking but stands briefly only to resume walking again, such an observation could acceptably be classified as standing. This could explain the misclassification between such classes.

In order to address the first issue, a fixed transition time between postures was used when extracting features; however, it could still not have been long enough for short postures. The second issue is harder to address as sometimes the posture itself may not be very well defined. It is advisable to utilise dedicated video commenting software for data labelling, allowing higher precision in the annotations.

# Chapter 6

# Conclusion

As discussed in Chapter 2, behavioural assessment methods are traditionally used for assessing trainee dogs' suitability for assistance work. In particular, ratings and test batteries are the most well-established tools, while observational tests are the least studied. More recently, canine activity monitors were identified as an emerging technology that could be used for monitoring assistance dogs. The prediction of future outcomes is considered the "holy grail" in this field of research, due to the potential for reducing the high costs of training.

Two primary objectives were set for this PhD research thesis as follows: (1) to explore the use of traditional and novel behavioural assessment methods for the prediction of future outcomes in trainee assistance dogs, and (2) to improve the performance of canine posture recognition systems for working dogs, by exploring the use of multiple sensors, data pre-processing techniques, and machine learning methods.

Accordingly, the present work contributed to the state-of-the-art in the field of behavioural assessment for assistance dogs, by validating a standardised rating for training outcome prediction for the first time; developing a new test battery for assistance dogs and validating its use for working outcome prediction; developing the most efficient classifiers for predicting future outcome in prospective assistance dogs using ratings and a test battery; and improving the performance of canine activity recognition systems by utilising inertial sensors in different positions, applying advanced feature extraction and selection techniques, and optimising the model architecture. The performance improvement achieved by the models developed for a working dog population is attributed to the different feature engineering methods implemented.

Section 6.1 compares the two different canine behavioural assessment methods investigated for the prediction of future outcomes in assistance dogs presented in Chapters 3 and 4. Section 6.2 describes the novelty

and contribution presented in each experimental chapter by summarising the gaps in the state-of-the-art addressed, highlighting the individual study goal, the methodology employed, the results achieved and finally answering the research questions associated with the study. Section 6.3 provides recommendations for researchers and practitioners.

## 6.1 Comparison between the Ratings and Test Battery

Chapters 3 and 4 aimed to predict future outcomes in assistance dogs. They examined the potential of two behavioural assessment methods to estimate the probability of withdrawal using machine learning techniques. This chapter provides a detailed comparison between the performance results obtained from the two behavioural assessment methods examined for the prediction of future outcomes.

Table 6.1 shows a detailed comparison of the performance metrics reported for each of the methods investigated for the prediction of training and working outcomes. The models built using C-BARQ, MCPQ-R and the ADTB at DC2 provided similar classification performance. The main difference was observed in PR, sensitivity, and f1-score, which were significantly better for DC2. This improvement reflects the fact that more withdrawn dogs were correctly classified in DC2 compared to all other models. A larger proportion of withdrawn dogs can be identified using the model built with data collected using the ADTB at ten weeks compared to all other models. Consequently, the DC2 model would allow dog training organisations to further minimise the cost of training by indicating a larger number of dogs that could be withdrawn.

Table 6.1: Performance metrics from the models created with data from the ratings C-BARQ and MCPQ-R and ADTB at DC1 and DC2.

| Dataset | ROC | PR | Precision | Sensitivity | Specificity | F1 |
|---|---|---|---|---|---|---|
| **Ratings** | | | | | | |
| C-BARQ | 0.84 | 0.64 | 0.57 | 0.36 | 0.94 | 0.44 |
| MCPQ-R | 0.85 | 0.58 | 0.75 | 0.33 | 0.97 | 0.46 |
| **Test Battery** | | | | | | |
| ADTB (DC1) | 0.74 | 0.37 | 0.50 | 0.24 | 0.94 | 0.32 |
| ADTB (DC2) | 0.84 | 0.74 | 0.67 | 0.67 | 0.92 | 0.67 |

Nonetheless, an important difference exists between those models. The models built using ratings predicted training outcomes while the ones built using the test batteries predicted working outcomes. Foyer et al. 2016 highlights the lack of consensus regarding the definition of a

successful assistance dog. Training outcome indicates whether the dog completed the training programme and was matched with a client; while working outcome denotes whether the dog completed the first year of the working programme. Additionally, L. Batt et al. 2010 suggests that working success is a more definite measure as it incorporates the effects of other important characteristics. Hence, even though the same performance was achieved, the DC2 model is considered superior and preferable because it predicts the probability of working outcomes instead of training outcomes with the same classification performance.

## 6.2 Concluding Summaries

An outline of each experimental chapter is provided below including details on the gap in the state-of-the-art addressed, the individual goals, the methodology employed, the results obtained and the answers to the research questions. Moreover, the novelty and contribution of the studies are highlighted.

### Chapter 3. Ratings

The most popular standardised dog behaviour rating reported in the literature for predicting training outcomes in assistance dogs is the Canine Behavioral Assessment and Research Questionnaire (C-BARQ). Its main downside is that it contains 100 context-based items, which makes it unsuitable for dog trainers to answer due to the number and nature of the questions. The goal of this chapter was to address this limitation by investigating the application of a shorter adjective-based rating; in particular, the Monash Canine Personality Questionnaire - Revised (MCPQ-R) was used for the first time for this same purpose.

Accordingly, two models were built separately using data collected from the same cohort using the C-BARQ (completed by puppy raisers at the start of formal training) and the MCPQ-R (completed by dog trainers at around ten weeks of training). Performance metrics of the two classifiers demonstrated that the MCPQ-R achieved similar results to the C-BARQ in the same cohort of trainee assistance dogs. Hence, this is the first study to reveal a relationship between MCPQ-R scores and training outcomes. Furthermore, the MCPQ-R was validated for predicting training outcomes in assistance dogs, as the performance results were comparable to the widely used C-BARQ. The novel C-BARQ model created outperformed all models presented in the literature, by employing a number of machine learning techniques for the first time in this area.

Conclusively, this chapter significantly contributes to the state-of-the-art by providing the most effective classifiers for predicting training outcomes in trainee assistance dogs, using data from two standardised

dog behaviour and personality questionnaires. The answers to the research questions investigated are presented below:

**Q.1** Is there a significant association between C-BARQ and MCPQ-R ratings and training outcomes?

Yes, significant associations ($p < 0.05$) with training outcomes were found in 29 out of the 114 C-BARQ items/factors and 13 out of the 26 MCPQ-R items. Univariate logistic regression was used to investigate the relationship between each of the variables individually and training outcomes (success or fail) in each dataset. This is the first study to demonstrate a relationship between MCPQ-R scores and training outcomes in assistance dogs.

**Q.2** Can statistical models estimate training outcomes based on C-BARQ and MCPQ-R ratings?

Yes, generalised linear models built using principal components calculated on features selected from the C-BARQ and MCPQ-R were validated for the prediction of training outcomes. Both models fit the data significantly better than the null model as indicated by the $\chi^2$ test. This was the first study to demonstrate that MCPQ-R scores can be used to predict training outcomes in assistance dogs.

**Q.3** Which is more predictive of training outcome: the C-BARQ or the MCPQ-R?

The models built on C-BARQ and MCPQ-R ratings provided comparable performance for the prediction of failure. They were compared using the area under the ROC curve as the main performance metric. C-BARQ achieved a ROC of 0.84 and MCPQ-R of 0.85, indicating a similar performance. Prediction results reveal that these new models outperformed previous studies using the C-BARQ.

## Chapter 4. Assistance Dog Test Battery

Test batteries have been utilised in the literature for the prediction of training outcomes in working dogs. No study was found to utilise test batteries to assess the behaviour of trainee assistance dogs at different stages of training, with the goal of predicting working outcomes. The goal of this chapter was to design and validate a new test battery with the purpose of estimating the suitability of trainee dogs for assistance work at different stages of the training programme. Hence, the novel Assistance Dog Test Battery (ADTB) was developed based on previous literature and evaluated for predicting working outcomes with machine learning models created based on behavioural measures recorded at three (DC1) and ten weeks (DC2) of formal training. Classification performance results demonstrated that while the DC1 model achieved

comparable results to the literature with a ROC of 0.74, the DC2 model outperformed previous studies on training outcome prediction using test batteries with a ROC of 0.84. Hence, this study demonstrated the relationship between the novel ADTB and working outcomes in assistance dogs. It also validated the use of novel machine learning models for predicting working outcomes at three and ten weeks of training. In addition to the behavioural measures recorded during the ADTB through the ethogram marking system, inertial data was monitored with three Actigraph GT9X Links, and physiological measures were also recorded with Polar H10 and Bioharness 3 devices. The author will make the data available through a public repository following the FAIR principles (Wilkinson et al. 2016), after the publication of this thesis and the corresponding research manuscript. This is the first test battery in trainee assistance working dogs to provide such a complete dataset, comprising not only behavioural but also cardio-respiratory measures.

Conclusively, this chapter adds to the body of knowledge by providing a new test battery for assessing the suitability of assistance dogs, which allowed the development of the most effective classifier for predicting working outcomes in trainee assistance dogs. As a result, organisations can reduce the costs of training by applying a similar methodology to assess the dog's suitability for assistance work. They are encouraged to add the database provided in the development of their own models.

The answers to the research questions investigated are presented below:

**Q.4** Are the new ethogram items and composite scores developed for evaluating the novel ADTB valid for assessing working outcomes?

Yes, a significant association ($p < 0.05$) with working outcomes was found in one ethogram item at DC1 ("Petting Confidence During"), and six ethogram items at DC2 during the following subtests: "walking" ("Pull strength" and "Pull on leash"), "tea towel" ("Second Response Indifferent"), "distractions" ("First Response Car"), "Kong" ("Interaction Response to stimulus"), "isolation" ("Response Time oriented" and "Response Exploration").

**Q.5** Can machine learning models predict working outcomes using ethogram items and composite scores for evaluating the novel ADTB?

Yes, the models developed using ethogram items and composite scores created to evaluate dog's behaviour during the Assistance Dog Test Battery (ADTB) subtest protocol delivered excellent prediction performance results. In particular, the DC1 model achieved a ROC of 0.74 matching previous studies on training outcome prediction using test batteries while the DC2 model achieved a ROC of 0.84, surpassing the state-of-the-art.

**Q.6** When is the optimal time to perform the novel ADTB on trainee dogs to maximise working outcome prediction performance?

This study investigated the predictive power of datasets built at two points in time: DC1 at three weeks, and DC2 at ten weeks, after the start of formal training. The results achieved by the two models trained independently on these datasets showed that the DC2 model had the best prediction performance as indicated by a 0.10 increase in the area under the ROC curve.

## Chapter 5. Posture Estimation

Posture recognition algorithms allow the automatic creation of ethograms (Gerencsér et al. 2013), as well as a range of other applications as discussed. However, these applications rely on the correct classification of postures. The goal of this study was to develop the first canine posture estimation system specifically for working dogs, with the purpose of improving classification performance. The system leveraged the fact that these canines wear a collar and harness to attach three IMUs (i.e., chest, back, neck). A novel supervised learning algorithm was developed to recognise static postures (i.e., standing, sitting, lying down) and dynamic activities (i.e. walking, body shake). For the first time in this field, advanced feature extraction techniques for time series data were employed. The novel algorithm utilised a feature selection method to choose the most important features for the prediction of the targeted postures. Moreover, different model architectures were utilised and a hyper-parameter optimisation technique was employed. The final optimal model contained features derived from the chest and back sensors only. It achieved an f1-weighted of 0.90. Comparisons with previous studies revealed the considerably higher performance of the model developed in this study. The importance of adding multiple IMUs to the harness (i.e., chest and back) in order to improve classification performance was demonstrated. Moreover, the employment of advanced machine learning techniques in the development of this novel model resulted in higher classification performance.

Conclusively, this work represents an advance in state of the art in posture estimation by expanding the current understanding of sensor position, feature extraction and feature importance as well as model architecture. It also provides an annotated dataset which is publicly available on Mendeley Data (link). The code developed in this study was also made publicly available on GitHub (link).

The answers to the research questions investigated are presented below:

**Q.7** Which IMU sensor position yields the most predictive features for canine posture recognition?

Features extracted from the back IMU were ranked higher in importance, followed by the chest IMU, and finally, the neck IMU was ranked a lot lower in importance.

**Q.8** Which feature extraction domain produces the most predictive features for canine posture recognition?

Features extracted using statistical domain techniques were the most important, followed by temporal and finally, spectral.

### Chapter 6. Conclusion

The answers to the research questions investigated in this chapter are presented below:

**Q.9** Which of the behavioural assessment methods are more predictive of future outcome - the rating (C-BARQ or MCPQ-R datasets) or ADTB ethograms (DC1 or DC2 datasets)?

The models built using C-BARQ, MCPQ-R and the ADTB at DC2 provided similar classification performance. However, the DC2 model predicted the probability of working outcome, unlike the others that only estimated training outcome. Moreover, the DC2 model was significantly better at identifying withdrawn dogs as indicated by the higher sensitivity. Hence, the DC2 model was the most predictive method of future outcome.

## 6.3 Recommendations and Future Work

This section provides recommendations for researchers and practitioners by highlighting the gaps in the state-of-the-art, which remain to be addressed and suggesting practical applications of this research. General recommendations about the use of behavioural assessment methods to predict future outcomes are discussed in the following paragraphs. Subsequently, specific recommendations considering each experimental chapter are made in the next subsections.

In terms of future outcome prediction, two types of behavioural assessments were investigated in this research: test battery and rating. The main advantage of the ADTB is that it does not require a person who knows the dog well to recall and volunteer the information, as in the case of C-BARQ or MCPQ-R. This simplifies the conditions for implementation and gives organisations more control to assess trainee dogs. Besides, test batteries are more standardised and less subjective than ratings, as they take place in a controlled environment and involve fewer raters.

More research is needed to identify the optimal timing for conducting suitability assessments in order to maximise reliability and minimise cost. The results from this study can be used as a reference for the prediction performance expected, considering the timing of the assessments. Nonetheless, the trade-off between the reliability and timing of predictions must be considered. Our results indicate that the further into training the assessments are conducted, the more reliable the predictions of future outcomes become; however, the cost of training also increases. Based on this, there is an opportunity for economic modelling, to identify the optimal assessment timing, based on training costs and prediction accuracy.

Additionally, the results achieved in this study can presumably be improved, by building models using data from a combination of behavioural assessment methods (i.e., ratings and test batteries), as shown in Bray et al. 2019. More advanced feature selection techniques (e.g., Random Forest importance) could be evaluated to further understand which behaviours improve the predictions of future outcomes. The use of diverse behavioural assessment methods would increase the complexity of data collection; however, the assessments can be simplified to include only elements used in the predictive models.

It is advisable for organisations to collect and manage their own data, in order to enable the development of predictive models specific to their practices and needs. This is particularly important due to the fact that organisations around the world reported different success rates, varying from 52% as in Bray et al. 2019 to 80% as in this sample. Using datasets with dissimilar class imbalances on the same model negatively affects classification performance (MacLean et al. 2018). Moreover, data sharing through public repositories is highly recommended in order to contribute to the advancement of the models and body of knowledge in this particular field.

In order to do that, assistance dog providers can follow a similar methodology to that presented in this study. Firstly, they should collect information on their trainee dogs such as behavioural (through standardised questionnaires), demographic (sex, breed, litter), training (start and end dates), outcome type (graduation or release) and status (assistance, guide, service, medical, etc). Secondly, they can develop their own models for training outcome prediction, by using their dataset to train models following the same methodology as that presented here. Finally, they are encouraged to anonymise and share the data publicly, in order to advance the current understanding of behaviour and training outcomes in assistance dogs.

Additionally, it is suggested to not only publish the research findings, but also make the full dataset openly available in data repositories for other researchers and practitioners following the FAIR data principles

(Wilkinson et al. 2016). These data represent a valuable asset which requires a significant investment of time and resources to attain. Open datasets in the area of early prediction of training/working outcomes in assistance dogs have the potential to substantially contribute to the advancement of the state-of-the-art.

Researchers are urged to publish the datasets from previous research on training outcomes with updated records on the working status one year after graduation. This was the second study to predict working outcomes (one year after graduation) using behaviour tests, while previous research only investigated the potential to predict training outcomes. In a cohort of 98 dogs in DC1 and 60 dogs in DC2, only one dog was withdrawn from the working programme for behavioural reasons. Nonetheless, this marginal difference might be specific to this cohort, or the IGDB, and might not reflect the reality of other assistance dog training organisations. Hence, future research in this area is encouraged.

## Chapter 3. Ratings

Researchers are encouraged to evaluate training outcome prediction with the MCPQ-R at an earlier stage. The earlier the data is collected, the higher the potential for predictions to create more efficiency in the training, as organisations can make informed decisions before more investment is made. In this study, the C-BARQ data were collected at the start of training, while the MCPQ-R data were collected at ten weeks of training. More research is needed to prove the total replacement of C-BARQ in favour of MCPQ-R at an earlier stage than ten weeks.

For instance, dog trainers could be invited to complete the MCPQ-R at around 3-4 weeks of training. Two main benefits are anticipated: (1) the prediction of training outcomes at an earlier stage allows issues to be addressed in the early days of training or starting the re-homing process; (2) the promotion of a higher engagement between trainer and dog from the start of training, as trainers are responsible for completing the questionnaire about the dog's personality. Another possibility is to invite puppy raisers to complete the MCPQ-R before the dogs start training, it is advisable to provide them with a description of the typical behaviours related to each of the adjectives.

In cases where C-BARQ data are not available at the beginning of training, organisations can obtain reliable predictions of training outcomes by using the MCPQ-R model at ten weeks of training. Finally, the validity of the MCPQ-R for training outcome prediction should also be evaluated in other types of working dogs. Researchers are encouraged to investigate the use of this personality rating for predicting training outcomes on trainee working dogs.

## Chapter 4. Assistance Dog Test Battery

The ADTB was designed to combine as many relevant subtests from previous research as possible; however, not all subtests produced variables significantly associated with working outcomes in trainee assistance dogs. In particular, no variables from the following five subtests were linked with working outcome at DC1 or DC2: "standing", "sitting", "lying down", "dog distraction", and "noise". Despite those results, it is recommended for researchers to perform the test in its entirety in other assistance dog training organisations with larger cohorts to investigate whether similar results are found before considering their removal.

Further investigation is recommended to evaluate the cost-benefit of earlier predictions, by assessing the trade-off between performance and timing. In this research, the working outcome prediction model derived from DC2 provided more reliable estimations than DC1. However, earlier predictions are preferable to reduce the costs of training and improve the welfare of trainee dogs. An optimal timing to perform the behaviour test for training outcome prediction might exist between three weeks and ten weeks of training.

In addition to the behavioural measures recorded during the ADTB through the ethogram marking system, inertial and physiological measures were also monitored. Inertial data were used to estimate posture as discussed in Chapter 5, but it was also identified as a potential predictor of future outcomes (training and/or working outcomes) (Mealin et al. 2020). Physiological measures were indicated to complement behavioural data as discussed in Chapter 2 and demonstrated to be a predictor of future outcome (Mealin et al. 2019).

The use of inertial and physiological data for training outcome prediction is outside the scope of the present work. However, the author will make the data available through a public repository following the FAIR principles (Wilkinson et al. 2016) after the publication of this thesis and a relevant research manuscript. A brief description of the datasets is provided below along with the type of research that could be executed and relevant references.

- Inertial data: use Inertial Measurement Unit (IMU) to predict future outcome (Mealin et al. 2020).
- Cardiac data: use RR Intervals (R-to-R) as recorded by the Polar H10 (Polar Electro Oy, Kempele, Finland) and Bioharness 3 (Zephyr Technology, Annapolis, MD, USA) devices, and/or ECG data from Bioharness 3 to predict future outcome (Mealin et al. 2019).
- Respiratory data: use breathing measures acquired by the Bioharness 3 to predict future outcomes. This has not been done before to my knowledge.

## Chapter 5. Posture Estimation

The review of the literature in the field of canine posture recognition, presented in Chapter 2, revealed that previous research failed to provide important methodological information. Similar to Chapter 5, researchers and practitioners are advised to report the following details in future research: data collection materials (sensor type and sampling rate); methods (number of dogs, breeds, context, types of behaviours, duration, repetitions); data pre-processing (window size and overlap); feature extraction (calculation, hyper-parameters); dataset statistics (size of training, development and validation/test sets); dataset splitting criteria (e.g. breed, subject, number of observations); dataset details (number of observations per behaviour); prediction algorithm (type, hyper-parameters); optimisation (hyper-parameters and search space) and evaluation (technique, hyper-parameters, metric). Performance reports should preferably include metrics for each class and confusion matrices should be reported.

In order to promote the interoperability of research outputs, researchers should make their datasets publicly available according to the FAIR principles (Wilkinson et al. 2016). Future research should estimate the performance of newly developed dog activity recognition models on previously published datasets whenever possible to benchmark results. Following the proposed recommendations will allow for easier comparison, reproducibility and knowledge transfer for the advancement of the state-of-the-art.

In terms of model development, different window sizes could be employed to improve the classification performance for classes shorter in duration, such as body shake. Accordingly, other techniques have been explored in the literature including KNN with Dynamic Time Warping (Kiyohara et al. 2015) and deep learning methods such as convolutional neural network (CNN) and long short-term memory (LSTM) motifs (Chambers et al. 2021).

Future research could investigate the potential of utilising posture estimation systems for predicting training outcomes in working dogs. Mealin et al. 2020 demonstrated that a neural network can produce comparable behavioural assessment scores to those produced by a human observer by using IMU measures in guide dogs. Similarly, researchers should explore the use of IMU measures for the direct prediction of training outcomes. Likewise, posture predictions could be used for the prediction of training outcomes.

## Additional Recommendations

Future research should investigate the relationship between cognition (how one interprets stimuli), affective state (how one feels about it) and

behavioural response (how one reacts to it). Traditionally, test battery protocols for canine assessment only monitored behavioural response through ethograms. The state-of-the-art can be expanded by evaluating not only behavioural responses but also affective state responses, as measured by physiological sensors (cardiac, respiratory and hormonal). The simultaneous monitoring of these diverse parameters will extend our understanding of the relationship between behaviour and physiological processes Bonne Beerda et al. 1997; Csoltova et al. 2017; Fallani et al. 2007; Gácsi et al. 2013; Kuhne et al. 2014; Maros et al. 2008; Part et al. 2014 as well as the contribution of health and personality traits-related factors in dogs.

# Bibliography


Adams, Jill U. (2008). “Genetics of Dog Breeding”. In: *Nature Education* 1.1, p. 44.

ADI (2022). *ADI Terms & Definitions*. https://assistancedogsinternational.org/resources/adi-terms-definitions/.

Aengwanich, Worapol et al. (2019). “Physiological Changes, Pain Stress, Oxidative Stress, and Total Antioxidant Capacity before, during, and after Castration in Male Dogs”. In: *Journal of Veterinary Behavior* 32, pp. 76–79. ISSN: 15587878. DOI: 10.1016/j.jveb.2019.04.004.

Agrafioti, F. et al. (2012). “ECG Pattern Analysis for Emotion Detection”. In: *IEEE Transactions on Affective Computing* 3.1, pp. 102–115. ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2011.28.

Aguilera, Ana M. et al. (2006). “Using Principal Components for Estimating Logistic Regression with High-Dimensional Multicollinear Data”. In: *Computational Statistics & Data Analysis* 50.8, pp. 1905–1924. ISSN: 01679473. DOI: 10.1016/j.csda.2005.03.011.

Akintola, Abimbola A. et al. (2016). “Comparative Analysis of the Equivital EQ02 Lifemonitor with Holter Ambulatory ECG Device for Continuous Measurement of ECG, Heart Rate, and Heart Rate Variability: A Validation Study for Precision and Accuracy”. In: *Frontiers in Physiology* 7. ISSN: 1664-042X. DOI: 10.3389/fphys.2016.00391.

Akiyoshi, Hideo et al. (2005). “Measurement of Plasma Chromogranin A Concentrations for Assessment of Stress Responses in Dogs with Insulin-Induced Hypoglycemia”. In: *American Journal of Veterinary Research* 66.10, pp. 1830–1835. ISSN: 0002-9645. DOI: 10.2460/ajvr.2005.66.1830.

Alberghina, Daniela et al. (2020). “Behavior Test for Seven-Week Old Puppies (Canis Familiaris): Inter-rater Reliability and Factors Associated with Test Performance”. In: *PLoS ONE* 15.7. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0236271.

Alcaidinho, Joelle et al. (2014). “Assessment of Working Dog Suitability from Quantimetric Data”. In: *NordiCHI’14*. Helsinki, Finland: Georgia Institute of Technology, pp. 1–3.

Antink, Christoph Hoog et al. (2019). “Sensor Fusion for Unobtrusive Respiratory Rate Estimation in Dogs”. In: *IEEE Sensors Journal* 19.16,

pp. 7072–7081. ISSN: 1530-437X, 1558-1748, 2379-9153. DOI: 10.1109/JSEN.2019.2912002.

Arata, Sayaka et al. (2010). "Important Behavioral Traits for Predicting Guide Dog Qualification". In: *Journal of Veterinary Medical Science* 72.5, pp. 539–545. ISSN: 0916-7250, 1347-7439. DOI: 10.1292/jvms.09-0512.

Arhant, Christine et al. (2020). "Balancing Skill against Difficulty - Behavior, Heart Rate and Heart Rate Variability of Shelter Dogs during Two Different Introductions of an Interactive Game". In: *Applied Animal Behaviour Science* 232, p. 105141. ISSN: 01681591. DOI: 10.1016/j.applanim.2020.105141.

Asher, Lucy et al. (2013). "A Standardized Behavior Test for Potential Guide Dog Puppies: Methods and Association with Subsequent Success in Guide Dog Training". In: *Journal of Veterinary Behavior* 8.6, pp. 431–438. ISSN: 15587878. DOI: 10.1016/j.jveb.2013.08.004.

Barandas, Marília et al. (2020). "TSFEL: Time Series Feature Extraction Library". In: *SoftwareX* 11, p. 100456. ISSN: 23527110. DOI: 10.1016/j.softx.2020.100456.

Barker, David A. et al. (2021). "Owner Reported Breathing Scores, Accelerometry and Sleep Disturbances in Brachycephalic and Control Dogs: A Pilot Study". In: *Veterinary Record* n/a.n/a, e135. ISSN: 2042-7670. DOI: 10.1002/vetr.135.

Barnard, Shanis et al. (2016). "Does Subjective Rating Reflect Behavioural Coding? Personality in 2 Month-Old Dog Puppies: An Open-Field Test and Adjective-Based Questionnaire". In: *PLOS ONE* 11.3. Ed. by Claudia Mettke-Hofmann, e0149831. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0149831.

Batt, Lara et al. (2008a). "Factors Associated with Success in Guide Dog Training". In: *Journal of Veterinary Behavior: Clinical Applications and Research* 3.4, pp. 143–151. ISSN: 15587878. DOI: 10.1016/j.jveb.2008.04.003.

Batt, Lara et al. (2008b). "The Effects of Structured Sessions for Juvenile Training and Socialization on Guide Dog Success and Puppy-Raiser Participation". In: *Journal of Veterinary Behavior* 3.5, pp. 199–206. ISSN: 1558-7878. DOI: 10.1016/j.jveb.2008.05.001.

– (2010). "Relationships between Puppy Management Practices and Reported Measures of Success in Guide Dog Training". In: *Journal of Veterinary Behavior: Clinical Applications and Research* 5.5, pp. 240–246. ISSN: 15587878. DOI: 10.1016/j.jveb.2010.02.004.

Batt, Lara S. et al. (2009). "The Value of Puppy Raisers' Assessments of Potential Guide Dogs' Behavioral Tendencies and Ability to Graduate". In: *Anthrozoös* 22.1, pp. 71–76. ISSN: 0892-7936, 1753-0377. DOI: 10.2752/175303708X390482.

Beauchaine, Theodore P (2015). "Respiratory Sinus Arrhythmia: A Transdiagnostic Biomarker of Emotion Dysregulation and

Psychopathology". In: *Current Opinion in Psychology* 3, pp. 43–47. ISSN: 2352250X. DOI: 10.1016/j.copsyc.2015.01.017.
Beerda, B et al. (2000). "Behavioural and Hormonal Indicators of Enduring Environmental Stress in Dogs". In: *Animal Welfare*, p. 15.
Beerda, Bonne et al. (1996). "The Use of Saliva Cortisol, Urinary Cortisol, and Catecholamine Measurements for a Noninvasive Assessment of Stress Responses in Dogs". In: *Hormones and Behavior* 30.3, pp. 272–279. ISSN: 0018506X. DOI: 10.1006/hbeh.1996.0033.
Beerda, Bonne et al. (1997). "Manifestations of Chronic and Acute Stress in Dogs". In: *Applied Animal Behaviour Science* 52.3-4, pp. 307–319. ISSN: 01681591. DOI: 10.1016/S0168-1591(96)01131-8.
Beerda, Bonne et al. (1998). "Behavioural, Saliva Cortisol and Heart Rate Responses to Different Types of Stimuli in Dogs". In: *Applied Animal Behaviour Science* 58, pp. 365–381.
Belda, B. et al. (2018). "Initial Evaluation of PetPace Activity Monitor". In: *The Veterinary Journal* 237, pp. 63–68. ISSN: 10900233. DOI: 10.1016/j.tvjl.2018.05.011.
Bergamasco, Luciana et al. (2010). "Heart Rate Variability and Saliva Cortisol Assessment in Shelter Dog: Human–Animal Interaction Effects". In: *Applied Animal Behaviour Science* 125, pp. 56–68. DOI: j.applanim.2010.03.002.
Bernston, Gary G. et al. (1997). "Heart Rate Variability: Origins, Methods, and Interpretive Caveats". In: *Psychophysiology* 34.6, pp. 623–648. DOI: 10.1111/j.1469-8986.1997.tb02140.x.
Berntson, Gary G. et al. (2008). "Cardiac Autonomic Balance versus Cardiac Regulatory Capacity". In: *Psychophysiology* 45.4, pp. 643–652. DOI: 10.1111/j.1469-8986.2008.00652.x.
Blackwell, Emily-Jayne et al. (2010). "Rapid Shaping of Behaviour Associated with High Urinary Cortisol in Domestic Dogs". In: *Applied Animal Behaviour Science* 124.3-4, pp. 113–120. ISSN: 01681591. DOI: 10.1016/j.applanim.2010.02.011.
Bodnariu, Alina (2008). "Indicators of Stress and Stress Assessment in Dogs". In: *Lucrari Stiinfice: Medicina Veterinara Timisoara (Scientifical Papers: Veterinary Medicine)*.
Bogaerts, Evelien et al. (2019). "Rejections in an Non-Purpose Bred Assistance Dog Population: Reasons, Consequences and Methods for Screening". In: *PLoS ONE* 14.6. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0218339.
Boiten, Frans A (1998). "The Effects of Emotional Behaviour on Components of the Respiratory Cycle". In: *Biological Psychology* 49.1-2, pp. 29–51. ISSN: 03010511. DOI: 10.1016/S0301-0511(98)00025-8.
Borell, Eberhard von et al. (2007). "Heart Rate Variability as a Measure of Autonomic Regulation of Cardiac Activity for Assessing Stress and Welfare in Farm Animals — A Review". In: *Physiology & Behavior* 92.3, pp. 293–316. DOI: 10.1016/j.physbeh.2007.01.007.

Bozkurt, Alper et al. (2014). “Toward Cyber-Enhanced Working Dogs for Search and Rescue”. In: *IEEE Intelligent Systems* 29.6, pp. 32–39. DOI: 10.1109/MIS.2014.77.

Brady, Karen et al. (2018). “A Systematic Review of the Reliability and Validity of Behavioural Tests Used to Assess Behavioural Characteristics Important in Working Dogs”. In: *Frontiers in Veterinary Science* 5. ISSN: 2297-1769. DOI: 10.3389/fvets.2018.00103.

Bray, Emily E. et al. (2017a). “Effects of Maternal Investment, Temperament, and Cognition on Guide Dog Success”. In: *Proceedings of the National Academy of Sciences* 114.34, pp. 9128–9133. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1704303114.

Bray, Emily E. et al. (2017b). “Temperament and Problem Solving in a Population of Adolescent Guide Dogs”. In: *Animal Cognition* 20.5, pp. 923–939. ISSN: 1435-9448, 1435-9456. DOI: 10.1007/s10071-017-1112-8.

Bray, Emily E. et al. (2019). “Predictive Models of Assistance Dog Training Outcomes Using the Canine Behavioral Assessment and Research Questionnaire and a Standardized Temperament Evaluation”. In: *Frontiers in Veterinary Science* 6, p. 49. ISSN: 2297-1769. DOI: 10.3389/fvets.2019.00049.

Bray, Emily E. et al. (2020). “Cognitive Characteristics of 8- to 10-Week-Old Assistance Dog Puppies”. In: *Animal Behaviour* 166, pp. 193–206. ISSN: 00033472. DOI: 10.1016/j.anbehav.2020.05.019.

– (2021a). “Dog Cognitive Development: A Longitudinal Study across the First 2 Years of Life”. In: *Animal Cognition* 24.2, pp. 311–328. ISSN: 1435-9448, 1435-9456. DOI: 10.1007/s10071-020-01443-7.

Bray, Emily E. et al. (2021b). “Early-Emerging and Highly Heritable Sensitivity to Human Communication in Dogs”. In: *Current Biology* 31.14, 3132–3136.e5. ISSN: 0960-9822. DOI: 10.1016/j.cub.2021.04.055.

Bray, Emily E. et al. (2021c). “Enhancing the Selection and Performance of Working Dogs”. In: *Frontiers in Veterinary Science* 8, p. 644431. ISSN: 2297-1769. DOI: 10.3389/fvets.2021.644431.

Briefer, Elodie F. et al. (2015a). “Emotions in Goats: Mapping Physiological, Behavioural and Vocal Profiles”. In: *Animal Behaviour* 99, pp. 131–143.

Briefer, Elodie F. et al. (2015b). “Segregation of Information about Emotional Arousal and Valence in Horse Whinnies”. In: *Scientific Reports*. DOI: 10.1038/srep09989.

Brodal, Per (2003). *The Central Nervous System: Structure and Function*. Oxford University Press. ISBN: 0-19-516560-8.

Brugarolas, Rita et al. (2012). “Posture Estimation for a Canine Machine Interface Based Training System”. In: *34th Annual International Conference of the IEEE EMBS*. IEEE, pp. 4489–4492.

Brugarolas, Rita et al. (2013a). “Behavior Recognition Based on Machine Learning Algorithms for a Wireless Canine Machine Interface”. In:

*IEEE International Conference on Body Sensor Networks (BSN)*. Cambridge, MA, USA: IEEE.

Brugarolas, Rita et al. (2013b). "Machine Learning Based Posture Estimation for a Wireless Canine Machine Interface". In: *2013 IEEE Topical Conference on Biomedical Wireless Technologies, Networks, and Sensing Systems*. IEEE, pp. 10–12. ISBN: 978-1-4673-2958-3 978-1-4673-2930-9. DOI: 10.1109/BioWireleSS.2013.6613658.

Brugarolas, Rita et al. (2014). "Wearable Wireless Biophotonic and Biopotential Sensors for Canine Health Monitoring". In: *IEEE Sensors*. Valencia, Spain: IEEE. DOI: 10.1109/ICSENS.2014.6985477.

Brugarolas, Rita et al. (2015). "Wearable Heart Rate Sensor Systems for Wireless Canine Health Monitoring". In: *IEEE Sensors Journal* 16.10, pp. 3454–3464. DOI: 10.1109/JSEN.2015.2485210.

Cannon, Walter Bradford (1916). *Bodily Changes in Pain, Hunger, Fear and Rage: An Account of Recent Researches into the Function of Emotional Excitement*. D. Appleton and Company.

Carballo, Fabricio et al. (2020). "Do Dogs Rescue Their Owners from a Stressful Situation? A Behavioral and Physiological Assessment". In: *Animal Cognition* 23.2, pp. 389–403. ISSN: 1435-9448, 1435-9456. DOI: 10.1007/s10071-019-01343-5.

Cardiosport (2021). *TP5+ Heart Rate Monitor, Cardiosport*. https://www.cardiosport.com/tp5-heart-rate-monitor.

Caron-Lormier, G. et al. (2016). "Using the Incidence and Impact of Health Conditions in Guide Dogs to Investigate Healthy Ageing in Working Dogs". In: *Veterinary Journal (London, England: 1997)* 207, pp. 124–130. ISSN: 1532-2971. DOI: 10.1016/j.tvjl.2015.10.046.

Caron-Lormier, Geoffrey et al. (2016). "Using the Incidence and Impact of Behavioural Conditions in Guide Dogs to Investigate Patterns in Undesirable Behaviour in Dogs". In: *Scientific Reports* 6.1, p. 23860. ISSN: 2045-2322. DOI: 10.1038/srep23860.

Carvalho, Inês et al. (2020). "The Combined Use of Salivary Cortisol Concentrations, Heart Rate, and Respiratory Rate for the Welfare Assessment of Dogs Involved in AAI Programs". In: *Journal of Veterinary Behavior* 36, pp. 26–33. ISSN: 15587878. DOI: 10.1016/j.jveb.2019.10.011.

Case, Linda P. (2005). *The Dog: Its Behavior, Nutrition, and Health*. 2nd ed. Ames, Iowa: Blackwell Pub. ISBN: 978-0-8138-1254-0.

Chambers, Robert D. et al. (2021). "Deep Learning Classification of Canine Behavior Using a Single Collar-Mounted Accelerometer: Real-World Validation". In: *Animals* 11.6, p. 1549. ISSN: 2076-2615. DOI: 10.3390/ani11061549.

Chen, Yanping et al. (2013). "DTW-D: Time Series Semi-Supervised Learning from a Single Example". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Chicago Illinois USA: ACM, pp. 383–391. ISBN: 978-1-4503-2174-7. DOI: 10.1145/2487575.2487633.

Chmelíková, E. et al. (2020). “Salivary Cortisol as a Marker of Acute Stress in Dogs: A Review”. In: *Domestic Animal Endocrinology* 72, p. 106428. ISSN: 07397240. DOI: 10.1016/j.domaniend.2019.106428.

Clay, Liam et al. (2020). “Comparison of Canine Behaviour Scored Using a Shelter Behaviour Assessment and an Owner Completed Questionnaire, C-BARQ”. In: *Animals* 10.10, p. 1797. ISSN: 2076-2615. DOI: 10.3390/ani10101797.

Cleghern, Zach et al. (2018). “Using Decision Tree Learning as an Interpretable Model for Predicting Candidate Guide Dog Success”. In.

Cobb, M.L. et al. (2016). “A Systematic Review and Meta-Analysis of Salivary Cortisol Measurement in Domestic Canines”. In: *Domestic Animal Endocrinology* 57, pp. 31–42. ISSN: 07397240. DOI: 10.1016/j.domaniend.2016.04.003.

Cobb, Mia et al. (2015). “The Advent of Canine Performance Science: Offering a Sustainable Future for Working Dogs”. In: *Behavioural Processes* 110, pp. 96–104. DOI: 10.1016/j.beproc.2014.10.012.

Colpoys, Jessica et al. (2021). “Evaluation of the FitBark Activity Monitor for Measuring Physical Activity in Dogs”. In: *Animals* 11.3, p. 781. DOI: 10.3390/ani11030781.

Cotur, Yasin et al. (2020). “Stretchable Composite Acoustic Transducer for Wearable Monitoring of Vital Signs”. In: *Advanced Functional Materials* 30.16, p. 1910288. ISSN: 1616-301X, 1616-3028. DOI: 10.1002/adfm.201910288.

Craig, Lydia (2016). “The Influence of Canine Aggression and Behavioral Treatment on Heart Rate Variability”. In: *Psychology Honors Projects* Paper 39, p. 81.

Craig, Lydia et al. (2017). “The Relationship between Heart Rate Variability and Canine Aggression”. In: *Applied Animal Behaviour Science* 188, pp. 59–67. DOI: 10.1016/j.applanim.2016.12.015.

Csoltova, Erika et al. (2017). “Behavioral and Physiological Reactions in Dogs to a Veterinary Examination: Owner-dog Interactions Improve Canine Well-Being”. In: *Physiology & Behavior* 177, pp. 270–281. ISSN: 00319384. DOI: 10.1016/j.physbeh.2017.05.013.

Davis, B.W. et al. (2004). “Assistance Dog Placement in the Pediatric Population: Benefits, Risks, and Recommendations for Future Application”. In: *Anthrozoös* 17.2, pp. 130–145. ISSN: 0892-7936. DOI: 10.2752/089279304786991765.

den Uijl, Ingrid et al. (2016). “Validation of a Collar-Mounted Accelerometer That Identifies Eight Canine Behavioural States, Including Those with Dermatologic Significance.” In: *Validation of a Collar-Mounted Accelerometer That Identifies Eight Canine Behavioural States, Including Those with Dermatologic Significance*. Weybridge, UK, pp. 81–84.

den Uijl, Ingrid et al. (2017). “External Validation of a Collar-Mounted Triaxial Accelerometer for Second-by-Second Monitoring of Eight

Behavioural States in Dogs". In: *PLOS ONE* 12.11. Ed. by Claire Wade, e0188481. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0188481.
Dollion, Nicolas et al. (2019). "Fear/Reactivity in Working Dogs: An Analysis of 37 Years of Behavioural Data from the Mira Foundation's Future Service Dogs". In: *Applied Animal Behaviour Science* 221, p. 104864. ISSN: 01681591. DOI: 10.1016/j.applanim.2019.104864.
Duffy, Deborah L. et al. (2008). "Behavioral Assessment of Guide and Service Dogs". In: *Journal of Veterinary Behavior* 3.4, pp. 186–188. ISSN: 15587878. DOI: 10.1016/j.jveb.2007.12.010.
– (2012). "Predictive Validity of a Method for Evaluating Temperament in Young Guide and Service Dogs". In: *Applied Animal Behaviour Science* 138.1, pp. 99–109. ISSN: 01681591. DOI: 10.1016/j.applanim.2012.02.011.
Engeland, W. C. et al. (1990). "Pituitary-Adrenal and Adrenomedullary Responses to Noise in Awake Dogs". In: *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 258.3, R672–R677. ISSN: 0363-6119, 1522-1490. DOI: 10.1152/ajpregu.1990.258.3.R672.
Ennik, Irma et al. (2006). "Suitability for Field Service in 4 Breeds of Guide Dogs". In: *Journal of Veterinary Behavior* 1.2, pp. 67–74. ISSN: 15587878. DOI: 10.1016/j.jveb.2006.06.004.
Equivital (2021). *Research Publications Using Equivital Products*. Tech. rep.
Essner, Ann et al. (2013). "Validity and Reliability of Polar® RS800CX Heart Rate Monitor, Measuring Heart Rate in Dogs during Standing Position and at Trot on a Treadmill". In: *Physiology & Behavior* 114–115, pp. 1–5. ISSN: 00319384. DOI: 10.1016/j.physbeh.2013.03.002.
Essner, Ann et al. (2015a). "Comparison of Polar® RS800CX Heart Rate Monitor and Electrocardiogram for Measuring Inter-Beat Intervals in Healthy Dogs". In: *Physiology & Behavior* 138, pp. 247–253. ISSN: 00319384. DOI: 10.1016/j.physbeh.2014.10.034.
Essner, Ann et al. (2015b). "Validity and Reliability Properties of Canine Short-Term Heart Rate Variability Measures—a Pilot Study". In: *Journal of Veterinary Behavior* 10.5, pp. 384–390. ISSN: 15587878. DOI: 10.1016/j.jveb.2015.05.006.
Evans, Katy et al. (2013). "Optimisation of the Breeding Programme for Guide Dogs in Th UK." In: DOI: 10.13140/2.1.2085.0886.
Evans, Katy M. et al. (2015). "Genetic Evaluation of Traits in a Standardized Behavioral Test for Potential Guide Dog Puppies Using Crossbreed Models". In: *Journal of Veterinary Behavior* 10.6, pp. 459–464. ISSN: 1558-7878. DOI: 10.1016/j.jveb.2015.08.004.
Fallani, Gaia et al. (2007). "Behavioral and Physiological Responses of Guide Dogs to a Situation of Emotional Distress". In: *Physiology & Behavior* 90, pp. 648–655. DOI: 10.1016/j.physbeh.2006.12.001.

Ferdinandy, Bence et al. (2020). “Challenges of Machine Learning Model Validation Using Correlated Behaviour Data: Evaluation of Cross-Validation Strategies and Accuracy Measures”. In: *PLOS ONE* 15.7. Ed. by Jishnu Das, e0236092. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0236092.

Flatt, Andrew A. et al. (2013). “Validity of the ithleteTM Smart Phone Application for Determining Ultra-Short-Term Heart Rate Variability”. In: *Journal of Human Kinetics* 39.1, pp. 85–92. ISSN: 1899-7562, 1640-5544. DOI: 10.2478/hukin-2013-0071.

Foster, Marc et al. (2020). “Inertial Measurement Based Heart and Respiration Rate Estimation of Dogs During Sleep for Welfare Monitoring”. In: *Proceedings of the Seventh International Conference on Animal-Computer Interaction*. Milton Keynes United Kingdom: ACM, pp. 1–6. ISBN: 978-1-4503-7574-0. DOI: 10.1145/3446002.3446125.

Foyer, Pernilla et al. (2014). “Behaviour and Experiences of Dogs during the First Year of Life Predict the Outcome in a Later Temperament Test”. In: *Applied Animal Behaviour Science* 155, pp. 93–100. ISSN: 01681591. DOI: 10.1016/j.applanim.2014.03.006.

Foyer, Pernilla et al. (2016). “Behavior and Cortisol Responses of Dogs Evaluated in a Standardized Temperament Test for Military Working Dogs”. In: *Journal of Veterinary Behavior* 11, pp. 7–12. DOI: 10.1016/j.jveb.2015.09.006.

Gácsi, Márta et al. (2013). “Human Analogue Safe Haven Effect of the Owner: Behavioural and Heart Rate Response to Stressful Social Stimuli in Dogs”. In: *PLoS ONE* 8.3. Ed. by Allan V. Kalueff, e58475. DOI: 10.1371/journal.pone.0058475.

Garmin et al. (2021). *Heart Rate Monitor HRM-Pro™, Garmin*. https://www.garmin.com/en-IE/p/682155.

Gerencsér, Linda et al. (2013). “Identification of Behaviour in Freely Moving Dogs (Canis Familiaris) Using Inertial Sensors”. In: *PLoS ONE* 8.10. Ed. by Gonzalo G. de Polavieja, e77814. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0077814.

Giles, David et al. (2015). “Validity of the Polar V800 Heart Rate Monitor to Measure RR Intervals at Rest”. In: *European Journal of Applied Physiology* 116.3, pp. 563–571. DOI: 10.1007/s00421-015-3303-9.

Gilgen-Ammann, Rahel et al. (2019). “RR Interval Signal Quality of a Heart Rate Monitor and an ECG Holter at Rest and during Exercise”. In: *European Journal of Applied Physiology* 119.7, pp. 1525–1532. ISSN: 1439-6319, 1439-6327. DOI: 10.1007/s00421-019-04142-5.

Goddard, M. E. et al. (1982). “Genetic and Environmental Factors Affecting the Suitability of Dogs as Guide Dogs for the Blind”. In: *Theoretical and Applied Genetics* 62.2, pp. 97–102. ISSN: 1432-2242. DOI: 10.1007/BF00293339.

– (1983). “Genetics of Traits Which Determine the Suitability of Dogs as Guide-Dogs for the Blind”. In: *Applied Animal Ethology* 9.3-4, pp. 299–315. ISSN: 03043762. DOI: 10.1016/0304-3762(83)90010-X.

Godoy, Lívea Dornela et al. (2018). "A Comprehensive Overview on Stress Neurobiology: Basic Concepts and Clinical Implications". In: *Frontiers in Behavioral Neuroscience* 12, p. 127. ISSN: 1662-5153. DOI: 10.3389/fnbeh.2018.00127.

Grossman, Paul et al. (2007). "Toward Understanding Respiratory Sinus Arrhythmia: Relations to Cardiac Vagal Tone, Evolution and Biobehavioral Functions". In: *Biological Psychology* 74.2, pp. 263–285. ISSN: 03010511. DOI: 10.1016/j.biopsycho.2005.11.014.

Hailstone, Jono et al. (2011). "Reliability and Validity of the Zephyr™ BioHarness™ to Measure Respiratory Responses to Exercise". In: *Measurement in Physical Education and Exercise Science* 15.4, pp. 293–300. DOI: 10.1080/1091367x.2011.615671.

Haixiang, Guo et al. (2017). "Learning from Class-Imbalanced Data: Review of Methods and Applications". In: *Expert Systems with Applications* 73, pp. 220–239. ISSN: 09574174. DOI: 10.1016/j.eswa.2016.12.035.

Harvey, Naomi D. (2021). *A Simple Guide to Inter-rater, Intra-rater and Test-retest Reliability for Animal Behaviour Studies*. Preprint. Open Science Framework. DOI: 10.31219/osf.io/8stpy.

Harvey, Naomi D. et al. (2016a). "Social Rearing Environment Influences Dog Behavioral Development". In: *Journal of Veterinary Behavior* 16, pp. 13–21. ISSN: 15587878. DOI: 10.1016/j.jveb.2016.03.004.

Harvey, Naomi D. et al. (2016b). "Test-Retest Reliability and Predictive Validity of a Juvenile Guide Dog Behavior Test". In: *Journal of Veterinary Behavior: Clinical Applications and Research* 11, pp. 65–76. DOI: 10.1016/j.jveb.2015.09.005.

Harvey, Naomi D. et al. (2017). "An Evidence-Based Decision Assistance Model for Predicting Training Outcome in Juvenile Guide Dogs". In: *PLOS ONE* 12.6. Ed. by Cheryl S. Rosenfeld, e0174261. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0174261.

Harvie, Helen et al. (2021). "Does Stress Run through the Leash? An Examination of Stress Transmission between Owners and Dogs during a Walk". In: *Animal Cognition* 24.2, pp. 239–250. ISSN: 1435-9448, 1435-9456. DOI: 10.1007/s10071-020-01460-6.

Hedhammar, Åke (2020). "Swedish Experiences From 60 Years of Screening and Breeding Programs for Hip Dysplasia—Research, Success, and Challenges". In: *Frontiers in Veterinary Science* 7, p. 228. ISSN: 2297-1769. DOI: 10.3389/fvets.2020.00228.

Hekman, Jessica et al. (2014). "Psychogenic Stress in Hospitalized Dogs: Cross Species Comparisons, Implications for Health Care, and the Challenges of Evaluation". In: *Animals* 4.2, pp. 331–347. ISSN: 2076-2615. DOI: 10.3390/ani4020331.

Henriksson, Josefine (2016). "Scores on Dog Personality Are Dependent on Questionnaire: A Comparison of Three Questionnaires". Bachelour. Linköping University.

Hiby, E et al. (2006). “Behavioural and Physiological Responses of Dogs Entering Re-Homing Kennels”. In: *Physiology & Behavior* 89.3, pp. 385–391. ISSN: 00319384. DOI: 10.1016/j.physbeh.2006.07.012.

Horváth, Zsuzsánna et al. (2007). “Three Different Coping Styles in Police Dogs Exposed to a Short-Term Challenge”. In: *Hormones and Behavior* 52.5, pp. 621–630. ISSN: 0018506X. DOI: 10.1016/j.yhbeh.2007.08.001.

Hosmer, D.W. et al. (2013). *Applied Logistic Regression*. Wiley Series in Probability and Statistics. Wiley. ISBN: 978-0-470-58247-3.

Hradecká, Lenka et al. (2015). “Heritability of Behavioural Traits in Domestic Dogs: A Meta-Analysis”. In: *Applied Animal Behaviour Science* 170, pp. 1–13. ISSN: 0168-1591. DOI: 10.1016/j.applanim.2015.06.006.

Hsu, Yuying et al. (2003). “Development and Validation of a Questionnaire for Measuring Behavior and Temperament Traits in Pet Dogs”. In: *Journal of the American Veterinary Medical Association* 223.9, pp. 1293–1300. ISSN: 0003-1488. DOI: 10.2460/javma.2003.223.1293.

Hunt, Rebecca L. et al. (2020). “Concurrent and Predictive Criterion Validity of a Puppy Behaviour Questionnaire for Predicting Training Outcome in Juvenile Guide Dogs”. In: *Animals* 10.12, p. 2382. ISSN: 2076-2615. DOI: 10.3390/ani10122382.

Hydbring-Sandberg, E (2004). “Physiological Reactions to Fear Provocation in Dogs”. In: *Journal of Endocrinology* 180.3, pp. 439–448. ISSN: 0022-0795, 1479-6805. DOI: 10.1677/joe.0.1800439.

ICHEC (2022). *Kay*. https://www.ichec.ie/about/infrastructure/kay.

IGDB (2018). *How Guide Dogs Are Trained - Irish Guide Dogs for the Blind*. https://guidedogs.ie/about-us/our-dog-training/.

– (2021). *Irish Guide Dogs for the Blind | Guide and Assistance Dogs*. https://www.guidedogs.ie/.

Invoxia (2021). *Invoxia Pet Tracker*. https://www.invoxia.com/ie/pet-tracker.

Ithlete (2021). *Bluetooth Smart and ANT+ Heart Rate Monitor - Myithlete*. https://www.myithlete.com/product/bluetooth-smart-ant-heart-rate-monitor/.

Ivković, Nedeljka et al. (2015). “Biomarkers of Stress in Saliva”. In: *Acta Facultatis Medicae Naissensis* 32.2, pp. 91–99. ISSN: 2217-2521. DOI: 10.1515/afmnai-2015-0010.

Jeong, Yi-Kyeong et al. (2020). “Evaluation of Salivary Vasopressin as an Acute Stress Biomarker in Healthy Dogs with Stress Due to Noise and Environmental Challenges”. In: *BMC Veterinary Research* 16.1, p. 331. ISSN: 1746-6148. DOI: 10.1186/s12917-020-02555-5.

Jezierski, Tadeusz et al. (2014). “Efficacy of Drug Detection by Fully-Trained Police Dogs Varies by Breed, Training Level, Type of Drug and Search Environment”. In: *Forensic Science International* 237,

pp. 112–118. ISSN: 03790738. DOI: 10.1016/j.forsciint.2014.01.013.
Jiang, Yue et al. (2008). "Techniques for Evaluating Fault Prediction Models". In: *Empirical Software Engineering* 13.5, pp. 561–595. ISSN: 1382-3256, 1573-7616. DOI: 10.1007/s10664-008-9079-3.
Johnstone, James A et al. (2012a). "BioharnessTM Multivariable Monitoring Device. Part I: Validity". In: p. 9.
– (2012b). "BioharnessTM Multivariable Monitoring Device. Part II: Reliability". In: p. 9.
Johnstone, James A et al. (2012c). "Field Based Reliability and Validity of the BioharnessTM Multivariable Monitoring Device". In: p. 11.
Jonckheer-Sheehy, Valerie S.M. et al. (2012). "Validation of a Polar® Human Heart Rate Monitor for Measuring Heart Rate and Heart Rate Variability in Adult Dogs under Stationary Conditions". In: *Journal of Veterinary Behavior: Clinical Applications and Research* 7.4, pp. 205–212. ISSN: 15587878. DOI: 10.1016/j.jveb.2011.10.006.
Jones, Amanda C. et al. (2005). "Temperament and Personality in Dogs (Canis Familiaris): A Review and Evaluation of Past Research". In: *Applied Animal Behaviour Science* 95.1-2, pp. 1–53. ISSN: 01681591. DOI: 10.1016/j.applanim.2005.04.008.
Jones, Amanda Claire (2008). "Development and Validation of a Dog Personality Questionnaire". In: *The University of Texas at Austin*.
Jones, Katherine E. et al. (2004). "Search-and-Rescue Dogs: An Overview for Veterinarians". In: *Journal of the American Veterinary Medical Association* 225.6, pp. 854–860. ISSN: 0003-1488. DOI: 10.2460/javma.2004.225.854.
Jones, Sarah et al. (2014). "Use of Accelerometers to Measure Stress Levels in Shelter Dogs". In: *Journal of Applied Animal Welfare Science* 17.1, pp. 18–28. ISSN: 1088-8705, 1532-7604. DOI: 10.1080/10888705.2014.856241.
Kanai, Kazutaka et al. (2008). "Circadian Variations in Salivary Chromogranin a Concentrations during a 24-Hour Period in Dogs". In: *Journal of Veterinary Science* 9.4, p. 421. ISSN: 1229-845X. DOI: 10.4142/jvs.2008.9.4.421.
Katayama, Maki et al. (2016). "Heart Rate Variability Predicts the Emotional State in Dogs". In: *Behavioural Processes* 128, pp. 108–112. DOI: 10.1016/j.beproc.2016.04.015.
Khan, Ali Mahmood (2011). "Personal State and Emotion Monitoring by Wearable Computing and Machine Learning". In: *Proceedings of HCI 2011 The 25th BCS Conference on Human Computer Interaction*. DOI: 10.14236/ewic/HCI2011.5.
Kikkawa, Aya et al. (2003). "Salivary Secretory IgA Concentrations in Beagle Dogs". In: *Journal of Veterinary Medical Science* 65.6, pp. 689–693. ISSN: 0916-7250, 1347-7439. DOI: 10.1292/jvms.65.689.
Kikkawa, Aya et al. (2005). "A Novel Method for Estimating the Adaptive Ability of Guide Dogs Using Salivary sIgA". In: *Journal of Veterinary*

*Medical Science* 67.7, pp. 707–712. ISSN: 0916-7250, 1347-7439. DOI: 10.1292/jvms.67.707.

Kim, J.-H. et al. (2013). "Measurement Accuracy of Heart Rate and Respiratory Rate during Graded Exercise and Sustained Exercise in the Heat Using the Zephyr BioHarness™." In: *Int J Sports Med.* 34(6), pp. 497–501. DOI: 10.1055/s-0032-1327661.

Kiyohara, Tatsuya et al. (2015). "Activity Recognition for Dogs Using Off-the-Shelf Accelerometer:" in: *Proceedings of the International Conference on Agents and Artificial Intelligence*. Lisbon, Portugal: SCITEPRESS - Science and and Technology Publications, pp. 100–110. ISBN: 978-989-758-073-4 978-989-758-074-1. DOI: 10.5220/0005212001000110.

Kleiger, Robert E. et al. (2005). "Heart Rate Variability: Measurement and Clinical Utility". In: *Annals of Noninvasive Electrocardiology* 10.1, pp. 88–101. DOI: 10.1111/j.1542-474x.2005.10101.x.

Kokocińska-Kusiak, Agata et al. (2021). "Canine Olfaction: Physiology, Behavior, and Possibilities for Practical Applications". In: *Animals* 11.8, p. 2463. ISSN: 2076-2615. DOI: 10.3390/ani11082463.

Kolevská, J. et al. (2003). "Circadian Rhythm of Cortisol Secretion in Dogs of Different Daily Activities". In: *Acta Veterinaria Brno* 72.4, pp. 599–605. ISSN: 0001-7213, 1801-7576. DOI: 10.2754/avb200372040599.

Koolhaas, J.M et al. (1999). "Coping Styles in Animals: Current Status in Behavior and Stress-Physiology". In: *Neuroscience & Biobehavioral Reviews* 23.7, pp. 925–935. ISSN: 01497634. DOI: 10.1016/S0149-7634(99)00026-3.

Koolhaas, Jaap M. et al. (2007). "Individual Variation in Coping with Stress: A Multidimensional Approach of Ultimate and Proximate Mechanisms". In: *Brain, Behavior and Evolution* 70.4, pp. 218–226. DOI: 10.1159/000105485.

Kooriyama, Takanori et al. (2021). "Salivary Stress Markers in Dogs: Potential Markers of Acute Stress". In: *Research in Veterinary Science* 141, pp. 48–55. ISSN: 00345288. DOI: 10.1016/j.rvsc.2021.10.009.

Kreibig, Sylvia D. (2010). "Autonomic Nervous System Activity in Emotion: A Review". In: *Biological Psychology* 84.3, pp. 394–421. DOI: 10.1016/j.biopsycho.2010.03.010.

Kuhne, Franziska et al. (2014). "Emotions in Dogs Being Petted by a Familiar or Unfamiliar Person: Validating Behavioural Indicators of Emotional States Using Heart Rate Variability". In: *Applied Animal Behaviour Science* 161, pp. 113–120. ISSN: 01681591. DOI: 10.1016/j.applanim.2014.09.020.

Kumpulainen, Pekka et al. (2018). "Dog Activity Classification with Movement Sensor Placed on the Collar". In: *Proceedings of the Fifth International Conference on Animal-Computer Interaction - ACI '18*. Atlanta, Georgia: ACM Press, pp. 1–6. ISBN: 978-1-4503-6219-1. DOI: 10.1145/3295598.3295602.

Kumpulainen, Pekka et al. (2021). “Dog Behaviour Classification with Movement Sensors Placed on the Harness and the Collar”. In: *Applied Animal Behaviour Science* 241, p. 105393. ISSN: 01681591. DOI: 10.1016/j.applanim.2021.105393.

Laborde, Sylvain et al. (2017). “Heart Rate Variability and Cardiac Vagal Tone in Psychophysiological Research – Recommendations for Experiment Planning, Data Analysis, and Data Reporting”. In: *Frontiers in Psychology* 08. DOI: 10.3389/fpsyg.2017.00213.

Ladha, Cassim et al. (2013). “Dog’s Life: Wearable Activity Recognition for Dogs”. In: *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp ’13*. ACM Press, p. 415. ISBN: 978-1-4503-1770-2. DOI: 10.1145/2493432.2493519.

Landis-Hanna, Amanda et al. (2016). “An Investigation of Ultra-Wideband Radar Technology to Evaluate Canine Heart Rate and Respiratory Rate”. In: *Journal of Veterinary Internal Medicine* 31. DOI: 10.1111/jvim.13963.

Lazarowski, Lucia et al. (2021). “Validation of a Behavior Test for Predicting Puppies’ Suitability as Detection Dogs”. In: *Animals* 11.4, p. 993. ISSN: 2076-2615. DOI: 10.3390/ani11040993.

Lee, H. S. et al. (2020). “Development of Novel Continuous and Interval Exercise Programs by Applying the FITT-VP Principle in Dogs”. In: *The Scientific World Journal* 2020, pp. 1–9. ISSN: 2356-6140, 1537-744X. DOI: 10.1155/2020/3029591.

Lee, Hae Sung et al. (2021). “Increased Total Serum Alkaline Phosphate Is Associated With Enhanced Bone Mineral Density in Dogs With Interval Exercise Training”. In: p. 17. DOI: 10.21203/rs.3.rs-146618/v1.

Leighton, E. A. (1997). “Genetics of Canine Hip Dysplasia”. In: *Journal of the American Veterinary Medical Association* 210.10, pp. 1474–1479. ISSN: 0003-1488.

Lensen, Catharina M.M. et al. (2015). “Saliva Sampling in Dogs: How to Select the Most Appropriate Procedure for Your Study”. In: *Journal of Veterinary Behavior* 10.6, pp. 504–512. ISSN: 15587878. DOI: 10.1016/j.jveb.2015.08.006.

Lensen, Rian C. M. M. et al. (2017). “Validity and Reliability of Cardiac Measures during Behavioural Tests in Pet Dogs at Home”. In: *Applied Animal Behaviour Science* 186, pp. 56–63. DOI: 10.1016/j.applanim.2016.10.011.

Lensen, Rian C. M. M. et al. (2019). “Physiological Stress Reactivity and Recovery Related to Behavioral Traits in Dogs (Canis Familiaris)”. In: *PLOS ONE* 14.9. Ed. by Christopher D. Lynn, e0222581. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0222581.

Levitt, Mairi (2013). “Perceptions of Nature, Nurture and Behaviour”. In: *Life Sciences, Society and Policy* 9. ISSN: 2195-7819. DOI: 10.1186/2195-7819-9-13.

Ley, Jacqui M. et al. (2008). "Personality Dimensions That Emerge in Companion Canines". In: *Applied Animal Behaviour Science* 110.3-4, pp. 305–317. ISSN: 01681591. DOI: 10.1016/j.applanim.2007.04.016.

Ley, Jacqui M. et al. (2009a). "A Refinement and Validation of the Monash Canine Personality Questionnaire (MCPQ)". In: *Applied Animal Behaviour Science* 116.2, pp. 220–227. ISSN: 01681591. DOI: 10.1016/j.applanim.2008.09.009.

Ley, Jacqui M. et al. (2009b). "Inter-Rater and Test–Retest Reliability of the Monash Canine Personality Questionnaire-Revised (MCPQ-R)". In: *Applied Animal Behaviour Science* 119.1, pp. 85–90. ISSN: 01681591. DOI: 10.1016/j.applanim.2009.02.027.

Liu, Yan et al. (2013). "Validity and Reliability of Multiparameter Physiological Measurements Recorded by the Equivital Lifemonitor During Activities of Various Intensities". In: *Journal of Occupational and Environmental Hygiene* 10.2, pp. 78–85. ISSN: 1545-9624, 1545-9632. DOI: 10.1080/15459624.2012.747404.

Lucidi, Pia et al. (2005). "Ethotest: A New Model to Identify (Shelter) Dogs' Skills as Service Animals or Adoptable Pets". In: *Applied Animal Behaviour Science* 95.1-2, pp. 103–122. ISSN: 01681591. DOI: 10.1016/j.applanim.2005.04.006.

Luque, Amalia et al. (2019). "The Impact of Class Imbalance in Classification Performance Metrics Based on the Binary Confusion Matrix". In: *Pattern Recognition* 91, pp. 216–231. ISSN: 00313203. DOI: 10.1016/j.patcog.2019.02.023.

Mackenzie, Stephen A. et al. (1985). "Heritability Estimate for Temperament Scores in German Shepherd Dogs and Its Genetic Correlation with Hip Dysplasia". In: *Behavior Genetics* 15.5, pp. 475–482. ISSN: 1573-3297. DOI: 10.1007/BF01066240.

MacLean, Evan L. et al. (2017). "Individual Differences in Cooperative Communicative Skills Are More Similar between Dogs and Humans than Chimpanzees". In: *Animal Behaviour* 126, pp. 41–51. ISSN: 0003-3472. DOI: 10.1016/j.anbehav.2017.01.005.

MacLean, Evan L. et al. (2018). "Enhanced Selection of Assistance and Explosive Detection Dogs Using Cognitive Measures". In: *Frontiers in Veterinary Science* 5, p. 236. ISSN: 2297-1769. DOI: 10.3389/fvets.2018.00236.

Mai, Dac et al. (2020). "Raising an Assistance Dog Puppy—Stakeholder Perspectives on What Helps and What Hinders". In: *Animals* 10.1, p. 128. DOI: 10.3390/ani10010128.

Mai, Dac L. et al. (2021). "Socialisation, Training, and Help-Seeking – Specific Puppy Raising Practices That Predict Desirable Behaviours in Trainee Assistance Dog Puppies". In: *Applied Animal Behaviour Science* 236, p. 105259. ISSN: 0168-1591. DOI: 10.1016/j.applanim.2021.105259.

Majikes, John et al. (2017). “Balancing Noise Sensitivity, Response Latency, and Posture Accuracy for a Computer-Assisted Canine Posture Training System”. In: *International Journal of Human-Computer Studies*.

Majikes, John J. et al. (2017). “Stimulus Control for Semi-autonomous Computer Canine-Training”. In: *Biomimetic and Biohybrid Systems*. Ed. by Michael Mangan et al. Vol. 10384. Cham: Springer International Publishing, pp. 279–290. ISBN: 978-3-319-63536-1 978-3-319-63537-8. DOI: 10.1007/978-3-319-63537-8_24.

Marcato, Marinara et al. (2022). “Assistance Dog Selection and Performance Assessment Methods Using Behavioural and Physiological Tools and Devices”. In: *Applied Animal Behaviour Science* 254, p. 105691. ISSN: 01681591. DOI: 10.1016/j.applanim.2022.105691.

Maros, Katalin et al. (2008). “Behavioural Correlation of Heart Rate Changes in Family Dogs”. In: *Applied Animal Behaviour Science* 109, pp. 329–341. DOI: 10.1016/j.applanim.2007.03.005.

McCraty, Rollin et al. (2009). “The Coherent Heart: Heart-Brain Interactions, Psychophysiological Coherence, and the Emergence of System-Wide Order”. In: *Integral Review* 5.2.

McCraty, Rollin et al. (2015). “Heart Rate Variability: New Perspectives on Physiological Mechanisms, Assessment of Self-regulatory Capacity, and Health Risk”. In: *Global Advances in Health and Medicine* 4.1, pp. 46–61. DOI: 10.7453/gahmj.2014.073.

Mealin, Sean (2021). “Sean_Mealin-PhD-Thesis.Pdf”. PhD thesis. North Carolina State University.

Mealin, Sean et al. (2019). “Predicting Guide Dog Temperament Evaluation Outcomes Using Raw ECG Signals”. In: *Proceedings of the Sixth International Conference on Animal-Computer Interaction*. Haifa Israel: ACM, pp. 1–12. ISBN: 978-1-4503-7693-8. DOI: 10.1145/3371049.3371053.

Mealin, Sean et al. (2020). “Using Inertial Measurement Unit Data for Objective Evaluations of Potential Guide Dogs”. In: *Proceedings of the Seventh International Conference on Animal-Computer Interaction*. Milton Keynes United Kingdom: ACM, pp. 1–11. ISBN: 978-1-4503-7574-0. DOI: 10.1145/3446002.3446064.

Menuge, Fanny et al. (2021). “Early Evaluation of Fearfulness in Future Guide Dogs for Blind People”. In: *Animals* 11.2, p. 412. ISSN: 2076-2615. DOI: 10.3390/ani11020412.

Meyer, Franziska et al. (2012). “Estimation of Genetic Parameters for Behavior Based on Results of German Shepherd Dogs in Switzerland”. In: *Applied Animal Behaviour Science* 140.1, pp. 53–61. ISSN: 0168-1591. DOI: 10.1016/j.applanim.2012.05.007.

Mirkó, Erika et al. (2012). “Preliminary Analysis of an Adjective-Based Dog Personality Questionnaire Developed to Measure Some Aspects of Personality in the Domestic Dog (Canis Familiaris)”. In: *Applied Animal*

*Behaviour Science* 138.1-2, pp. 88–98. ISSN: 01681591. DOI: 10.1016/j.applanim.2012.02.016.

Mirkó, Erika et al. (2013). "Association between Subjective Rating and Behaviour Coding and the Role of Experience in Making Video Assessments on the Personality of the Domestic Dog (Canis Familiaris)". In: *Applied Animal Behaviour Science* 149.1-4, pp. 45–54. ISSN: 01681591. DOI: 10.1016/j.applanim.2013.10.003.

Mizukoshi, Mina et al. (2008). "Evaluation of the Potential Suitability of Guide Dog Candidates by Continuous Observation during Training". In: *Journal of Veterinary Behavior* 3.5, pp. 193–198. ISSN: 1558-7878. DOI: 10.1016/j.jveb.2008.05.002.

Muggeridge, David Joseph et al. (2021). "Measurement of Heart Rate Using the Polar OH1 and Fitbit Charge 3 Wearable Devices in Healthy Adults During Light, Moderate, Vigorous, and Sprint-Based Exercise: Validation Study". In: *JMIR mHealth and uHealth* 9.3, e25313. ISSN: 2291-5222. DOI: 10.2196/25313.

Müller, Andre Matthias et al. (2019). "Heart Rate Measures From Wrist-Worn Activity Trackers in a Laboratory and Free-Living Setting: Validation Study". In: *JMIR mHealth and uHealth* 7.10, e14120. ISSN: 2291-5222. DOI: 10.2196/14120.

Murphy, Julie A. (1998). "Describing Categories of Temperament in Potential Guide Dogs for the Blind". In: *Applied Animal Behaviour Science* 58.1-2, pp. 163–178. DOI: 10.1016/s0168-1591(97)00047-6.

Nakahara, Eri et al. (2016). "Canine Emotional States Assessment with Heart Rate Variability". In: *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, pp. 1–3. DOI: 10.1109/apsipa.2016.7820868.

Nepi, Daniele et al. (2016). "Validation of the Heart:Rate Signal Provided by the Zephyr BioHarness 3.0". In: *2016 Computing in Cardiology Conference*. DOI: 10.22489/CinC.2016.106-358.

Oberbauer, A. M. et al. (2017). "Long-Term Genetic Selection Reduced Prevalence of Hip and Elbow Dysplasia in 60 Dog Breeds". In: *PLoS ONE* 12.2, e0172918. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0172918.

Olstad, Bjørn Harald et al. (2020). "Validation of the Polar OH1 and M600 Optical Heart Rate Sensors during Front Crawl Swim Training". In: *PLOS ONE* 15.4. Ed. by Dalton Müller Pessôa Filho, e0231522. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0231522.

Orozco, Sonia C. et al. (2021). "Efficacy of High-intensity Interval Training Compared with Moderate-intensity Continuous Training on Maximal Aerobic Potency in Dogs: Trial Protocol for a Randomised Controlled Clinical Study". In: *Veterinary Record Open* 8.1. ISSN: 2399-2050, 2052-6113. DOI: 10.1002/vro2.4.

Ortmeyer, Heidi et al. (2018). "Combining Actigraph Link and PetPace Collar Data to Measure Activity, Proximity, and Physiological

Responses in Freely Moving Dogs in a Natural Environment". In: *Animals* 8.12, p. 230. ISSN: 2076-2615. DOI: 10.3390/ani8120230.
Overall, Karen (2013). *Manual of Clinical Behavioral Medicine for Dogs and Cats - E-Book*. Elsevier Health Sciences. ISBN: 978-0-323-24065-9.
Overall, Karen L. et al. (2014). "Genetics and Behavior: A Guide for Practitioners". In: *The Veterinary Clinics of North America. Small Animal Practice* 44.3, pp. 483–505. ISSN: 1878-1306. DOI: 10.1016/j.cvsm.2014.01.006.
Pagani, M. et al. (1986). "Power Spectral Analysis of Heart Rate and Arterial Pressure Variabilities as a Marker of Sympatho-Vagal Interaction in Man and Conscious Dog". In: *Circulation Research* 59.2, pp. 178–193. DOI: 10.1161/01.res.59.2.178.
Palestrini, Clara et al. (2005). "Heart Rate and Behavioural Responses of Dogs in the Ainsworth's Strange Situation: A Pilot Study". In: *Applied Animal Behaviour Science* 94.1-2, pp. 75–88. DOI: 10.1016/j.applanim.2005.02.005.
Parenti, Lindsay et al. (2015). "Selecting Quality Service Dogs". In: *The APDT chronicle of the dog* 2015.summer, pp. 71–77.
Part, C. E. et al. (2014). "Physiological, Physical and Behavioural Changes in Dogs (Canis Familiaris) When Kennelled: Testing the Validity of Stress Parameters". In: *Physiology & Behavior* 133, pp. 260–271. DOI: 10.1016/j.physbeh.2014.05.018.
PetPace (2021). *Smart Dog Collar - Heart Rate Monitor & More | Petpace*. https://petpace.com/.
Petry, N. M. et al. (2015). "Perceptions about Professionally and Non-Professionally Trained Hypoglycemia Detection Dogs". In: *Diabetes research and clinical practice* 109.2, pp. 389–396. ISSN: 0168-8227. DOI: 10.1016/j.diabres.2015.05.023.
Pirrone, Federica et al. (2019). "Salivary Vasopressin as a Potential Non–Invasive Biomarker of Anxiety in Dogs Diagnosed with Separation–Related Problems". In: *Animals* 9.12, p. 1033. ISSN: 2076-2615. DOI: 10.3390/ani9121033.
Polar (2017). *The New Polar H10 Heart Rate Sensor Is Now Available*. https://www.polar.com/blog/new-polar-h10-heart-rate-sensor-2017/.
– (2019). *Polar H10 Heart Rate Sensor System*. White Paper. Polar, p. 6.
– (2021). *Polar H10 | Heart rate monitor chest strap*. https://www.polar.com/ie-en/products/accessories/polar_h10_heart_rate_sensor.
Polgár, Zita et al. (2019). "Assessing the Welfare of Kennelled Dogs—A Review of Animal-Based Measures". In: *Applied Animal Behaviour Science* 213, pp. 1–13. ISSN: 01681591. DOI: 10.1016/j.applanim.2019.02.013.
Porges, Stephen W. (2007). "The Polyvagal Perspective". In: *Biological Psychology* 74.2, pp. 116–143. ISSN: 03010511. DOI: 10.1016/j.biopsycho.2006.06.009.

Posluns, Julie A. et al. (2017). "Comparing Two Canine Personality Assessments: Convergence of the MCPQ-R and DPQ and Consensus between Dog Owners and Dog Walkers". In: *Applied Animal Behaviour Science* 188, pp. 68–76. ISSN: 01681591. DOI: 10.1016/j.applanim.2016.12.013.

Quintana, Daniel S. et al. (2014). "Considerations in the Assessment of Heart Rate Variability in Biobehavioral Research". In: *Frontiers in Psychology* 5. DOI: 10.3389/fpsyg.2014.00805.

Rayment, Diana J. et al. (2015). "Applied Personality Assessment in Domestic Dogs: Limitations and Caveats". In: *Applied Animal Behaviour Science* 163, pp. 1–18. DOI: 10.1016/j.applanim.2014.11.020.

Rayment, Diana J. et al. (2016). "Investigating Canine Personality Structure Using Owner Questionnaires Measuring Pet Dog Behaviour and Personality". In: *Applied Animal Behaviour Science* 180, pp. 100–106. ISSN: 01681591. DOI: 10.1016/j.applanim.2016.04.002.

Reynard, Alison et al. (2011). "Heart Rate Variability as a Marker of Self-Regulation". In: *Applied Psychophysiology and Biofeedback* 36.3, pp. 209–215. DOI: 10.1007/s10484-011-9162-1.

Ribeiro, Cristina et al. (2008). "Wireless Estimation of Canine Pose for Search and Rescue". In: *2008 IEEE International Conference on System of Systems Engineering*. IEEE, pp. 1–6. ISBN: 978-1-4244-2172-5. DOI: 10.1109/SYSOSE.2008.4724172.

Ribeiro, Cristina et al. (2009). "Canine Pose Estimation: A Computing for Public Safety Solution". In: *2009 Canadian Conference on Computer and Robot Vision*. IEEE, pp. 37–44. ISBN: 978-1-4244-4211-9 978-0-7695-3651-4. DOI: 10.1109/CRV.2009.38.

Robins, R.W. et al. (2009). *Handbook of Research Methods in Personality Psychology*. Guilford Publications. ISBN: 978-1-60623-656-7.

Rodriguez, Kerri E. et al. (2020). "The Effects of Assistance Dogs on Psychosocial Health and Wellbeing: A Systematic Literature Review". In: *PLOS ONE* 15.12. Ed. by Geilson Lima Santana, e0243302. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0243302.

Ryan, Morag G. et al. (2019). "Physiological Indicators of Attachment in Domestic Dogs (Canis Familiaris) and Their Owners in the Strange Situation Test". In: *Frontiers in Behavioral Neuroscience* 13, p. 162. ISSN: 1662-5153. DOI: 10.3389/fnbeh.2019.00162.

Sandri, Misa et al. (2015). "Salivary Cortisol Concentration in Healthy Dogs Is Affected by Size, Sex, and Housing Context". In: *Journal of Veterinary Behavior* 10.4, pp. 302–306. ISSN: 15587878. DOI: 10.1016/j.jveb.2015.03.011.

Schwab-Richards, Rachel et al. (2014). "Use of activity monitors for assessment of pruritus in an acute model of canine atopic dermatitis". In: *Veterinary Dermatology* 25.5, 441–e69. ISSN: 1365-3164. DOI: 10.1111/vde.12133.

Serpell, James (2021). *C-BARQ: Canine Behavioral Assessment & Research Questionnaire*. https://vetapps.vet.upenn.edu/cbarq/.

Serpell, James et al. (2005). "Effects of Breed, Sex, and Neuter Status on Trainability in Dogs". In: *Anthrozoos A Multidisciplinary Journal of The Interactions of People & Animals* 18, pp. 196–207. DOI: 10.2752/089279305785594135.

Serpell, James A. et al. (2001). "Development and Validation of a Novel Method for Evaluating Behavior and Temperament in Guide Dogs". In: *Applied Animal Behaviour Science* 72.4, pp. 347–364. DOI: 10.1016/s0168-1591(00)00210-0.

Serpell, James A. et al. (2016). "Aspects of Juvenile and Adolescent Environment Predict Aggression and Fear in 12-Month-Old Guide Dogs". In: *Frontiers in Veterinary Science* 3. ISSN: 2297-1769. DOI: 10.3389/fvets.2016.00049.

Sforzini, E. et al. (2009). "Evaluation of Young and Adult Dogs' Reactivity". In: *Journal of Veterinary Behavior* 4.1, pp. 3–10. ISSN: 15587878. DOI: 10.1016/j.jveb.2008.09.035.

Shaffer, Fred et al. (2017). "An Overview of Heart Rate Variability Metrics and Norms". In: *Frontiers in Public Health* 5, p. 258. ISSN: 2296-2565. DOI: 10.3389/fpubh.2017.00258.

Shouldice, Victoria L. et al. (2019). "Expression of Behavioural Traits in Goldendoodles and Labradoodles". In: *Animals : an Open Access Journal from MDPI* 9.12. ISSN: 2076-2615. DOI: 10.3390/ani9121162.

Sinn, David L. et al. (2010). "Personality and Performance in Military Working Dogs: Reliability and Predictive Validity of Behavioral Tests". In: *Applied Animal Behaviour Science* 127.1-2, pp. 51–65. DOI: 10.1016/j.applanim.2010.08.007.

Siracusa, C. et al. (2010). "A Comparison between a Direct and an Indirect Method for Assessing Aggressiveness in Dogs". In: *Journal of Veterinary Behavior* 5.1, p. 55. ISSN: 15587878. DOI: 10.1016/j.jveb.2009.09.016.

Skandakumar, S et al. (1995). "Salivary IgA: A Possible Stress Marker In Dogs". In: *Animal Welfare* 4.4, pp. 339–350.

Slabbert, J.M et al. (1999). "Early Prediction of Adult Police Dog Efficiency—a Longitudinal Study". In: *Applied Animal Behaviour Science* 64.4, pp. 269–288. ISSN: 01681591. DOI: 10.1016/S0168-1591(99)00038-6.

Sokolova, Marina et al. (2009). "A Systematic Analysis of Performance Measures for Classification Tasks". In: *Information Processing & Management* 45.4, pp. 427–437. ISSN: 03064573. DOI: 10.1016/j.ipm.2009.03.002.

Sousa, Antonio A R et al. (2017). "Evaluation of Cortisol Levels in Female Dogs Undergoing Ovariosalpingohysterectomy after Autohemoteraphy". In: 6.3, p. 3.

Souza, Carla Caroline Franzini de et al. (2017). "Autonomic, Endocrine and Behavioural Responses to Thunder in Laboratory and Companion

Dogs". In: *Physiology & Behavior* 169, pp. 208–215. ISSN: 00319384. DOI: 10.1016/j.physbeh.2016.12.006.

Srithunyarat, Thanikul et al. (2016). "Catestatin, Vasostatin, Cortisol, Temperature, Heart Rate, Respiratory Rate, Scores of the Short Form of the Glasgow Composite Measure Pain Scale and Visual Analog Scale for Stress and Pain Behavior in Dogs before and after Ovariohysterectomy". In: *BMC Research Notes* 9.1, p. 381. ISSN: 1756-0500. DOI: 10.1186/s13104-016-2193-1.

Starling, Melissa et al. (2013). "Conceptualising the Impact of Arousal and Affective State on Training Outcomes of Operant Conditioning". In: *Animals* 3.2, pp. 300–317. ISSN: 2076-2615. DOI: 10.3390/ani3020300.

Stellato, Anastasia C. et al. (2019). "Effect of High Levels of Background Noise on Dog Responses to a Routine Physical Examination in a Veterinary Setting". In: *Applied Animal Behaviour Science* 214, pp. 64–71. ISSN: 01681591. DOI: 10.1016/j.applanim.2019.03.009.

Stephen, Jacqueline M. et al. (2006). "A Longitudinal Evaluation of Urinary Cortisol in Kennelled Dogs, Canis Familiaris". In: *Physiology & Behavior* 87.5, pp. 911–916. ISSN: 00319384. DOI: 10.1016/j.physbeh.2006.02.015.

Suunto (2021). *Smart Heart Rate Belt Suunto*. https://www.suunto.com/en-ie/Products/Heart-Rate-Belts/suunto-smart-heart-rate-belt/.

Svartberg, Kenth (2002). "Shyness–Boldness Predicts Performance in Working Dogs". In: *Applied Animal Behaviour Science* 79.2, pp. 157–174. ISSN: 01681591. DOI: 10.1016/S0168-1591(02)00120-X.

– (2005). "A Comparison of Behaviour in Test and in Everyday Life: Evidence of Three Consistent Boldness-Related Personality Traits in Dogs". In: *Applied Animal Behaviour Science* 91.1-2, pp. 103–128. ISSN: 01681591. DOI: 10.1016/j.applanim.2004.08.030.

Svartberg, Kenth et al. (2002). "Personality Traits in the Domestic Dog (Canis Familiaris)". In: *Applied Animal Behaviour Science* 79.2, pp. 133–155. ISSN: 01681591. DOI: 10.1016/S0168-1591(02)00121-1.

Svobodová, Ivona et al. (2008). "Testing German Shepherd Puppies to Assess Their Chances of Certification". In: *Applied Animal Behaviour Science* 113.1, pp. 139–149. ISSN: 0168-1591. DOI: 10.1016/j.applanim.2007.09.010.

Svobodová, Ivona et al. (2014). "Cortisol and Secretory Immunoglobulin A Response to Stress in German Shepherd Dogs". In: *PLoS ONE* 9.3. Ed. by Jan S. Suchodolski, e90820. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0090820.

Taylor, Katy D. et al. (2006). "The Development and Assessment of Temperament Tests for Adult Companion Dogs". In: *Journal of Veterinary Behavior* 1.3, pp. 94–108. ISSN: 15587878. DOI: 10.1016/j.jveb.2006.09.002.

Team, R Core (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Thayer, Julian F. et al. (2009). "Heart Rate Variability, Prefrontal Neural Function, and Cognitive Performance: The Neurovisceral Integration Perspective on Self-regulation, Adaptation, and Health". In: *Annals of Behavioral Medicine* 37.2, pp. 141–153. DOI: 10.1007/s12160-009-9101-z.

Tomkins, L. et al. (2010). "Sensory Jump Test as a Measure of Sensory (Visual) Lateralization in Dogs (Canis Familiaris)". In: *Journal of Veterinary Behavior Clinical Applications and Research* 5, pp. 256–267. DOI: 10.1016/j.jveb.2010.02.005.

Tomkins, Lisa M. et al. (2011). "Behavioral and Physiological Predictors of Guide Dog Success". In: *Journal of Veterinary Behavior: Clinical Applications and Research* 6.3, pp. 178–187. DOI: 10.1016/j.jveb.2010.12.002.

– (2012). "Associations between Motor, Sensory and Structural Lateralisation and Guide Dog Success". In: *The Veterinary Journal* 192.3, pp. 359–367. ISSN: 10900233. DOI: 10.1016/j.tvjl.2011.09.010.

Turner, Anne I. et al. (2020). "Psychological Stress Reactivity and Future Health and Disease Outcomes: A Systematic Review of Prospective Evidence". In: *Psychoneuroendocrinology* 114, p. 104599. ISSN: 03064530. DOI: 10.1016/j.psyneuen.2020.104599.

van der Waaij, E. H. et al. (2008). "Genetic Analysis of Results of a Swedish Behavior Test on German Shepherd Dogs and Labrador Retrievers1". In: *Journal of Animal Science* 86.11, pp. 2853–2861. ISSN: 0021-8812. DOI: 10.2527/jas.2007-0616.

van Rooy, Diane et al. (2019). "Behavioral Genetics of Dog Breeds". In: *Encyclopedia of Animal Behavior (Second Edition)*. Ed. by Jae Chun Choe. Oxford: Academic Press, pp. 312–322. ISBN: 978-0-12-813252-4. DOI: 10.1016/B978-0-12-809633-8.90059-1.

Virtanen, Juhani et al. (2018). "Evaluation of Dry Electrodes in Canine Heart Rate Monitoring". In: *Sensors* 18.6, p. 1757. ISSN: 1424-8220. DOI: 10.3390/s18061757.

Vollmer-Conna, Ute et al. (2020). "The Well-Being of Medical Students: A Biopsychosocial Approach". In: *Australian & New Zealand Journal of Psychiatry* 54.10, pp. 997–1006. ISSN: 0004-8674, 1440-1614. DOI: 10.1177/0004867420924086.

Wahoo (2021). *TICKR X Heart Rate Monitor, Wahoo Fitness EU*. https://eu.wahoofitness.com/devices/heart-rate-monitors/tickr-x/buy.

Weippert, Matthias et al. (2010). "Comparison of Three Mobile Devices for Measuring R–R Intervals and Heart Rate Variability: Polar S810i, Suunto T6 and an Ambulatory ECG System". In: *European Journal of Applied Physiology* 109.4, pp. 779–786. ISSN: 1439-6319, 1439-6327. DOI: 10.1007/s00421-010-1415-9.

Weiss, Emily (2002). “Selecting Shelter Dogs for Service Dog Training”. In: *Journal of Applied Animal Welfare Science* 5.1, pp. 43–62. ISSN: 1088-8705, 1532-7604. DOI: 10.1207/S15327604JAWS0501_4.

Weiss, Emily et al. (1997). “Service Dog Selection Tests: Effectiveness for Dogs from Animal Shelters”. In: *Applied Animal Behaviour Science* 53.4, pp. 297–308. ISSN: 01681591. DOI: 10.1016/S0168-1591(96)01176-8.

Wells, Deborah L. (2021). “Paw Preference as a Tool for Assessing Emotional Functioning and Welfare in Dogs and Cats: A Review”. In: *Applied Animal Behaviour Science* 236, p. 105148. ISSN: 01681591. DOI: 10.1016/j.applanim.2020.105148.

Wells, Deborah L. et al. (2016). “Comparing Lateral Bias in Dogs and Humans Using the Kong™ Ball Test”. In: *Applied Animal Behaviour Science* 176, pp. 70–76. ISSN: 01681591. DOI: 10.1016/j.applanim.2016.01.010.

Whistle (2021). *Whistle FIT*. https://www.whistle.com/products/whistle-fit-pet-health-tracker-activity-monitor.

Wiener, Pamela et al. (2016). “Use of Questionnaire-Based Data to Assess Dog Personality”. In: *Journal of Veterinary Behavior* 16, pp. 81–85. ISSN: 15587878. DOI: 10.1016/j.jveb.2016.10.007.

Wiggett, Cindy (2006). “Guide Dog Ownership and Psychological Well-Being /”. In: p. 324.

Wilkinson, Mark D. et al. (2016). “The FAIR Guiding Principles for Scientific Data Management and Stewardship”. In: *Scientific Data* 3.1, p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18.

Wilsson, Erik et al. (1997a). “The Use of a Behaviour Test for Selection of Dogs for Service and Breeding. II. Heritability for Tested Parameters and Effect of Selection Based on Service Dog Characteristics”. In: *Applied Animal Behaviour Science* 54.2, pp. 235–241. ISSN: 0168-1591. DOI: 10.1016/S0168-1591(96)01175-6.

– (1997b). “The Use of a Behaviour Test for the Selection of Dogs for Service and Breeding, I: Method of Testing and Evaluating Test Results in the Adult Dog, Demands on Different Kinds of Service Dogs, Sex and Breed Differences”. In: *Applied Animal Behaviour Science* 53.4, pp. 279–295. DOI: 10.1016/s0168-1591(96)01174-4.

– (1998). “Behaviour Test for Eight-Week Old Puppies—Heritabilities of Tested Behaviour Traits and Its Correspondence to Later Behaviour”. In: *Applied Animal Behaviour Science* 58.1-2, pp. 151–162. ISSN: 01681591. DOI: 10.1016/S0168-1591(97)00093-2.

Wilsson, Erik et al. (2012). “Are There Differences between Behavioral Measurement Methods? A Comparison of the Predictive Validity of Two Ratings Methods in a Working Dog Program”. In: *Applied Animal Behaviour Science* 141.3, pp. 158–172. ISSN: 01681591. DOI: 10.1016/j.applanim.2012.08.012.

Wirth, Kathleen E. et al. (2008). “The Economic Costs and Benefits of Dog Guides for the Blind”. In: *Ophthalmic Epidemiology* 15.2, pp. 92–98. DOI: 10.1080/09286580801939353.

Wood, Cassidy (2020). “Assessing the Role Personality Plays in Puppy Raisers and Guide Dogs in Training”. In: *Honors Theses*.

Wormald, Dennis et al. (2016). “Physiological Stress Coping and Anxiety in Greyhounds Displaying Inter-Dog Aggression”. In: *Applied Animal Behaviour Science* 180, pp. 93–99. DOI: 10.1016/j.applanim.2016.04.007.

– (2017). “Reduced Heart Rate Variability in Pet Dogs Affected by Anxiety-Related Behaviour Problems”. In: *Physiology & Behavior* 168, pp. 122–127. DOI: 10.1016/j.physbeh.2016.11.003.

Wright, Hannah F et al. (2011). “Development and Validation of a Psychometric Tool for Assessing Impulsivity in the Domestic Dog (Canis Familiaris)”. In: p. 17.

– (2012). “Behavioural and Physiological Correlates of Impulsivity in the Domestic Dog (Canis Familiaris)”. In: *Physiology & Behavior* 105.3, pp. 676–682. ISSN: 00319384. DOI: 10.1016/j.physbeh.2011.09.019.

Yashari, Jonathan M. et al. (2015). “Evaluation of a Novel Canine Activity Monitor for At-Home Physical Activity Analysis”. In: *BMC Veterinary Research*. DOI: 10.1186/s12917-015-0457-y.

Zephyr (2021). *Bioharness, Zephyr Performance Systems*. https://www.zephyranywhere.com/system/components.

Zilocchi, Marcella et al. (2016). “Effects of Physical Activity on Dog Behavior”. In: *Dogbehavior* 2. DOI: 10.4454/db.v2i2.34.

Zingaro, Margherita et al. (2018). “Are the Livestock Guarding Dogs Where They Are Supposed to Be?” In: *Applied Animal Behaviour Science* 198, pp. 89–94. ISSN: 0168-1591. DOI: 10.1016/j.applanim.2017.10.002.

Ziv, Gal (2017). “The Effects of Using Aversive Training Methods in Dogs—A Review”. In: *Journal of Veterinary Behavior: Clinical Applications and Research* 19, pp. 50–60. DOI: 10.1016/j.jveb.2017.02.004.

Zon, E. van (2017). *Assistance Dogs for Elderly with Middle and Late Stage Dementia : A Study on the Implementation of Assistance Dogs in the Care of Elderly with Middle and Late Stage Dementia Who Live in Healthcare Facilities*. https://essay.utwente.nl/73857/. Info:Eu-Repo/Semantics/masterThesis.

Zupan, Manja et al. (2016). “Assessing Positive Emotional States in Dogs Using Heart Rate and Heart Rate Variability”. In: *Physiology & Behavior* 155, pp. 102–111. DOI: 10.1016/j.physbeh.2015.11.027.

# Acronyms

**ACC** Accelerometer.

**ADTB** Assistance Dog Test Battery.

**AEEC** Animal Experimentation Ethics Committee.

**BM** Belgian Malinois.

**BR** Breathing Rate.

**C-BARQ** Canine Behavioral Assessment and Research Questionnaire.

**CALIN** Celtic Advanced Life Science Innovation Network.

**CI** Confidence Interval.

**DC1** Data Collection 1.

**DC2** Data Collection 2.

**ECG** Electrocardiography.

**GLM** Generalised Linear Model.

**GR** Golden Retriever.

**GYR** Gyroscope.

**HR** Heart Rate.

**HRV** Heart Rate Variability.

**IBI** Interbeat Interval.

**ICHEC** Irish Centre for High-End Computing.

**IGDB** Irish Guide Dogs for the Blind.

**IMU** Inertial Measurement Unit.

**IMUs** Inertial Measurement Units.

**KK** Kai Ken.

**LOOCV** Leave One Out Cross Validation.
**LR** Labrador Retriever.
**LRxGR** Labrador Retriever crossed with Golden Retriever.

**M** Mean.
**MAG** Magnetometer.
**MCPQ-R** Monash Canine Personality Questionnaire - Revised.

**NPV** Negative Predictive Value.

**OR** Odds Ratio.

**PCA** Principal Component Analysis.
**PPV** Positive Predictive Value.
**PR** Precision-Recall.

**R-to-R** RR Intervals.
**RFE** Recursive Feature Elimination.
**ROC** Receiver Operating Characteristic.

**SD** Standard Deviation.
**SREC** Social Research Ethics Committee.
**SU** Swansea University.

**TNI** Tyndall National Institute.
**TNR** True Negative Rate.
**TPR** True Positive Rate.

**UCC** University College Cork.

# Appendix A

# Physiological Assessment - Hormonal

This section reviews three hormonal biomarkers; namely, cortisol, salivary Immunoglobulin A (sIgA), catecholamines, and chromogranin A (CgA). They are discussed in the context of evaluating canine emotional responses to stimuli, as well as a means of assessing their cognitive abilities and personality characteristics. These aspects are particularly relevant when estimating the suitability of prospective dogs for assistance work.

Cortisol has been widely used as a physiological marker of stress in dogs (M.L. Cobb et al. 2016). It has been employed in research as a means of measuring physiological response to several stressors (Bonne Beerda et al. 1998; Souza et al. 2017) and fear-eliciting stimuli in pet dogs (Hydbring-Sandberg 2004) and guide dogs (Menuge et al. 2021). Previous research with dogs living in various housing conditions has revealed cortisol as a good welfare indicator (Beerda et al. 2000; Part et al. 2014; Stephen et al. 2006; Bergamasco et al. 2010).

However, some contradictory results have been reported regarding the relationships between cortisol levels and training outcome, problem-solving ability and learning performance in dogs. Cortisol levels were higher in trainee military dogs approved for further training than in non-approved dogs both before and after taking a temperament test (Foyer et al. 2016). Nonetheless, cortisol levels from trainee guide dogs collected before and after a temperament test did not show a significant association with subsequent training success (L. Batt et al. 2008a). Higher cortisol concentration was associated with reduced proficiency at a problem-solving task, and fearful and anxious behaviour (Bray et al. 2017b), but in another study, it resulted in an enhanced ability to learn (Blackwell et al. 2010). Coping styles have also been reported to affect the HPA axis response, in particular, reactive dogs show a lower level of HPA axis activation compared to proactive

dogs who exhibit a higher activation (Blackwell et al. 2010; Horváth et al. 2007).

Cortisol levels can be assayed to assess short-term stress response using blood plasma or saliva (Bonne Beerda et al. 1996) and long-term welfare indicator using urine, faeces or hair (Chmelíková et al. 2020). Plasma, saliva and urine cortisol levels reflect acute stressors (Chmelíková et al. 2020), on the other hand, hair and faeces levels result from accrued welfare conditions (Polgár et al. 2019).

The assessment of cortisol in saliva presents many advantages over other methods. C. M. Lensen et al. 2015 discusses considerations for selecting the most appropriate saliva sampling procedure for cortisol analysis. Plasma collection is more invasive and stressful to the animal while urine samples require creatinine to also be assayed as cortisol levels are expressed in relation to creatinine levels (cortisol:creatinine ratios) to adjust for dilution of urine.

M.L. Cobb et al. 2016 conducted a systematic review and meta-analysis of salivary cortisol measurements in dogs. After inspecting 31 raw datasets, several significant effects were found to influence cortisol levels; namely, sex, neuter status, age, regular living environment (Beerda et al. 2000), time in the environment before testing, testing environment, owner presence during testing, and collection media (Sandri et al. 2015). Chmelíková et al. 2020 reviews current literature on salivary cortisol studies on dogs and proposes that the circadian rhythms (Kolevská et al. 2003) and coping strategies (J. M. Koolhaas et al. 2007; Horváth et al. 2007; Blackwell et al. 2010; Rayment et al. 2015) also affect cortisol response. Kooriyama et al. 2021 highlights the high inter-individual and intra-individual variability reported by M.L. Cobb et al. 2016 and Chmelíková et al. 2020. Studies involving cortisol should be carefully designed and conducted to control for the intensity of the stressors and appropriate timing of sample collection (Kooriyama et al. 2021).

Salivary secretory Immunoglobulin A (sIgA) concentration levels have been considered a potential biomarker of well-being, in particular, decreased levels of sIgA suggest higher levels of stress in adult dogs.Skandakumar et al. 1995 reported a significant negative correlation between cortisol concentrations and sIgA levels in dog saliva; however, Svobodová et al. 2014 did not find any significant relationship between cortisol and sIgA differences after stress. High sIgA levels were associated with high adaptive ability in police dogs (Skandakumar et al. 1995) and guide dogs (Kikkawa et al. 2005). Similarly, R. C. M. M. Lensen et al. 2019 found that high sIgA measures were associated with desirable traits such as high levels of trainability and low levels of both non-social fear and touch sensitivity. Kikkawa et al. 2005 demonstrated that sIgA concentration was significantly

higher in acceptable than in rejected dogs in a guide dog programme.

Catecholamines levels reach the peak at 1 min and remain elevated for approximately 3 min (Engeland et al. 1990). This rapid response poses challenges as catecholamines have low concentrations and instability in saliva samples (Ivković et al. 2015). The monitoring of chromogranin A (CgA) offers a practical alternative as it is more stable than catecholamines and is co-released with adrenaline (Kooriyama et al. 2021). In dogs, plasma levels of CgA increased after physiological stress (insulin-induced hypoglycemia) (Akiyoshi et al. 2005) and salivary levels positively correlated with the duration of behavioural signs of stress (Ryan et al. 2019). CgA does not appear to be influenced by circadian rhythms (Kanai et al. 2008); however, baseline levels of salivary concentration reported were quite different Ryan et al. 2019.

R. C. M. M. Lensen et al. 2019 conducted a behaviour test on pet dogs as puppies (17 weeks) and then as adults (60 weeks) to investigate the association between behaviour as assessed by C-BARQ and physiological reactivity (10min post-test) and recovery (40min post-test) for three salivary biomarkers: cortisol, CgA and sIgA. Several associations were found to exist between physiological biomarkers and behavioural traits at the same test stage as well as physiological variables at T1 and behaviour at T2. The predictive validity of a puppy's physiological stress response to estimate behavioural traits from C-BARQ was inconclusive. Contradictory relationships were found between physiological parameters and behavioural traits at different points in time, suggesting that developmental factors play an important role.

Kooriyama et al. 2021 reviews alternative salivary stress biomarkers in dogs; namely CgA, sIgA, oxytocin (OT), and vasopressin (AVP). AVP and OT have opposing effects in regulating emotional behaviours. Contradictory results have been presented regarding AVP concentration in response to stressors. Pirrone et al. 2019 found AVP concentrations were significantly higher in dogs affected by separation problems than normal dogs immediately after the end of the separation period, no significant difference was found in pre- and post-stress concentrations of AVP. On the other hand, Jeong et al. 2020 reported a significant decrease in AVP concentration between pre- and post-stress. No significant changes in OT were detected in different groups or phases Pirrone et al. 2019.

Results from published research indicate that hormonal markers offer a potential method for assessing a dog's suitability for assistance work either directly (Kikkawa et al. 2005) or indirectly through the evaluation of important behavioural traits (R. C. M. M. Lensen et al. 2019) that contribute to their success (Ryan et al. 2019) or as estimation of stress level perceived by the prospective dog in a given situation (Akiyoshi et al. 2005).

# Appendix B

# Behavioural Assessment - Ratings

## B.1 Canine Behavioral Assessment and Research Questionnaire (C-BARQ)

This questionnaire was developed, standardised and validated by J. A. Serpell et al. 2001; Hsu et al. 2003; Duffy et al. 2012. The rating scales are divided into sections and all items had an NA option which was used to indicate that the respondent had never observed the dog in the situation described.

Table B.1: Detailed description of the factors and the original items used to calculate them.

| Factor | Description | Items |
|---|---|---|
| F_01 Trainability | Dog shows a willingness to attend to the owner and obey simple commands. Dog is not easily distracted, tends to be a fast learner, responds positively to correction, and will fetch or retrieve objects | 1, 2, 3, 4, 5, 6, 7, 8 |
| F_02 Stranger-directed aggression | Dog shows threatening or aggressive responses to strangers approaching or invading the dog's or the owner's personal space, territory, or home range | 10, 11, 12, 15, 16, 18, 20, 21, 22, 28 |
| F_03 Owner-directed aggression | Dog shows threatening or aggressive responses to the owner or other members of the household when challenged, manhandled, stared at, stepped over, or when approached while in possession of food or objects | 9, 13, 14, 17, 19, 25, 30, 31 |

| | | |
|---|---|---|
| F_04 Dog rivalry | Dog shows aggressive or threatening responses to other familiar dogs in the same household | 32, 33, 34, 35 |
| F_05 Stranger-directed fear | Dog shows fearful or wary responses when approached directly by strangers | 36, 37, 39, 40 |
| F_06 Nonsocial fear | Dog shows fearful or wary responses to sudden or loud noises (e.g. thunder), traffic, and unfamiliar objects and situations | 38, 41, 42, 44, 47, 48 |
| F_07 Dog-directed aggression | Dog shows threatening or aggressive responses when approached directly by unfamiliar dogs | 23, 24, 26, 29 |
| F_08 Dog-directed fear | Dog shows fearful or wary responses when approached directly by unfamiliar dogs | 45, 46, 52, 53 |
| F_09 Touch sensitivity | Dog shows fearful or wary responses to potentially painful or uncomfortable procedures, including bathing, grooming, nail clipping, and veterinary examinations | 43, 49, 50, 51 |
| F_10 Separation-related behaviour | Dog vocalises and/or is destructive when separated from the owner, often accompanied or preceded by behavioural and autonomic signs of anxiety including restlessness, loss of appetite, trembling, and excessive salivation | 54, 55, 56, 57, 58, 59, 60, 61 |
| F_11 Excitability | Dog displays strong reaction to potentially exciting or arousing events, such as going for walks or car trips, doorbells, the arrival of visitors, and the owner arriving home; has difficulty calming down after such events | 62, 63, 64, 65, 66, 67 |
| F_12 Attachment and attention-seeking | Dog maintains close proximity to the owner or other members of the household, solicits affection or attention, and displays agitation when the owner gives attention to third parties | 68, 69, 70, 71, 72, 73 |
| F_13 Chasing | Dog chases cats, birds, and/or other small animals, given the opportunity | 27, 74, 75, 76 |
| F_14 Energy level | Dog is energetic, “always on the go”, and/or playful | 91, 92 |

Table B.2: C-BARQ rating sections and items.

| **Section 1. Trainability** |
|---|
| 0 = Never, 1 = Seldom, 2 = Sometimes, 3 = Usually, 4 = Always |

| |
|---|
| I_1 When off the leash returns immediately when called |
| I_2 Obeys the sit command immediately |
| I_3 Obeys the stay command immediately |
| I_4 Seems to attend listen closely to everything you say or do |
| I_5 Slow to respond to correction or punishment thick skinned |
| I_6 Slow to learn new tricks or tasks |
| I_7 Easily distracted by interesting sights sounds or smells |
| I_8 Will fetch or attempt to fetch sticks balls or objects |
| **Section 2. Aggression** |
| 0 = No aggression (no visible signs of aggression) |
| 1-3 = Mild to moderate (growling, barking, baring teeth) |
| 4 = Serious aggression (snaps bites or attempts to bite) |
| I_9 When verbally corrected or punished scolded shouted at etc by you or a household member |
| I_10 When approached directly by an unfamiliar adult while being walked/exercised on a leash |
| I_11 When approached directly by an unfamiliar child while being walked/exercised on a leash |
| I_12 Toward unfamiliar persons approaching the dog while she is in your car at the gas station for example |
| I_13 When toys bones or other objects are taken away by a household member |
| I_14 When bathed or groomed by a household member |
| I_15 When an unfamiliar person approaches you or another member of your family at home |
| I_16 When unfamiliar persons approach you or another member of your family away from your home |
| I_17 When approached directly by a household member while she the dog is eating |
| I_18 When mailmen or other delivery workers approach your home |
| I_19 When his/her food is taken away by a household member |
| I_20 When strangers walk past your home while your dog is outside or in the yard |
| I_21 When an unfamiliar person tries to touch or pet the dog |
| I_22 When joggers cyclists roller bladers or skate boarders pass your home |
| I_23 When approached directly by an unfamiliar male dog while being walked/exercised on a leash |
| I_24 When approached directly by an unfamiliar female dog while being walked/exercised on a leash |
| I_25 When stared at directly by a member of the household |
| I_26 Toward unfamiliar dogs visiting your home |
| I_27 Toward cats squirrels or other small animals entering your yard |
| I_28 Toward unfamiliar persons visiting your home |
| I_29 When barked growled or lunged at by another unfamiliar dog |
| I_30 When stepped over by a member of the household |

I_31 When you or a household member retrieves food or objects stolen by the dog
I_32 Towards another familiar dog in your household NA if no other dogs
I_33 When approached at a favourite resting/sleeping place by another familiar household dog NA if no other dogs
I_34 When approached while eating by another familiar household dog NA if no other dogs
I_35 When approached while playing with chewing a favourite toy bone object etc by another familiar household dog NA if no other dogs

**Section 3. Fear and Anxiety**
0 = No fear/anxiety (no visible signs of fear/anxiety)
1-3 = Mild to moderate
4 = Extreme fear (cowers, retreats or hides)

I_36 When approached directly by an unfamiliar adult while away from your home
I_37 When approached directly by an unfamiliar child while away from your home
I_38 In response to sudden or loud noises eg vacuum cleaner car backfire road drills objects being dropped etc
I_39 When unfamiliar persons visit your home
I_40 When an unfamiliar person tries to touch or pet the dog
I_41 In heavy traffic
I_42 In response to strange or unfamiliar objects on or near the sidewalk eg plastic trash bags leaves litter flags flapping etc
I_43 When examined/treated by a veterinarian
I_44 During thunderstorms firework displays or similar events
I_45 When approached directly by an unfamiliar dog of the same or larger size
I_46 When approached directly by an unfamiliar dog of a smaller size
I_47 When first exposed to unfamiliar situations eg first car trip first time in elevator first visit to veterinarian etc
I_48 In response to wind or windblown objects
I_49 When having nails clipped by a household member
I_50 When groomed or bathed by a household member
I_51 When having his/her feet towed by a member of the household
I_52 When unfamiliar dogs visit your home
I_53 When barked growled or lunged at by an unfamiliar dog

**Section 4. Separation-related behavior**
0 = Never, 1 = Seldom, 2 = Sometimes, 3 = Usually, 4 = Always

I_54 Shaking shivering or trembling
I_55 Excessive salivation
I_56 Restlessness agitation pacing
I_57 Whining
I_58 Howling
I_59 Chewing scratching at doors floor windows curtains etc
I_60 Loss of appetite

| |
|---|
| I_61 Barking |
| **Section 5. Excitability** |
| 0 = Calm (little or no special reaction)<br>1-3 = Mild to moderate<br>4 = Extremely excitable (over-reacts, hard to calm down) |
| I_62 When you or other members of the household come home after a brief absence |
| I_63 When playing with you or other members of your household |
| I_64 When doorbell rings |
| I_65 Just before being taken for a walk |
| I_66 Just before being taken on a car trip |
| I_67 When visitors arrive at your home |
| **Section 6. Attachment and Attention-seeking** |
| 0 = Never, 1 = Seldom, 2 = Sometimes, 3 = Usually, 4 = Always |
| I_68 Displays a strong attachment for one particular member of the household |
| I_69 Tends to follow you or other members of household about the house from room to room |
| I_70 Tends to sit close to or in contact with you or others when you are sitting down |
| I_71 Tends to nudge nuzzle or paw you or others for attention when you are sitting down |
| I_72 Becomes agitated whines jumps up tries to intervene when you or others show affection for another person |
| I_73 Becomes agitated whines jumps up tries to intervene when you show affection for another dog or animal |
| **Section 7. Miscellaneous** |
| 0 = Never, 1 = Seldom, 2 = Sometimes, 3 = Usually, 4 = Always |
| I_74 Chases or would chase cats given the opportunity |
| I_75 Chases or would chase birds given the opportunity |
| I_76 Chases or would chase squirrels rabbits and other small animals given the opportunity |
| I_77 Escapes or would escape from home or yard given the chance |
| I_78 Rolls in animal droppings or other smelly substances |
| I_79 Eats own or other animals droppings or faeces |
| I_80 Chews inappropriate objects |
| I_81 Mounts objects furniture or people |
| I_82 Begs persistently for food when people are eating |
| I_83 Steals food |
| I_84 Nervous or frightened on stairs |
| I_85 Pulls excessively hard when on the leash |
| I_86 Urinates against objects furnishings in your home |
| I_87 Urinates when approached petted handled or picked up |
| I_88 Urinates when left alone at night or during the daytime |
| I_89 Defecates when left alone at night or during the daytime |
| I_90 Hyperactive restless has trouble settling down |

I_91 Playful puppyish boisterous
I_92 Active energetic always on the go
I_93 Stares intently at nothing visible
I_94 Snaps at invisible flies
I_95 Chases own tailwind end
I_96 Chases/follows shadows light spots etc
I_97 Barks persistently when alarmed or excited
I_98 Licks him/herself excessively
I_99 Licks people or objects excessively
I_100 Displays other bizarre strange or repetitive behaviours

### B.1.1 Descriptive Statistics

Table B.3: C-BARQ rating descriptive statistics including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing values for all items and factors.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| **Section 1. Trainability (0=Never, 4=Always)** | | | | | | | |
| I_01 When off the leash returns immediately when called | 1.00 | 3.00 | 3.00 | 3.11 | 4.00 | 4.00 | 0 (0.00) |
| I_02 Obeys the sit command immediately | 2.00 | 3.00 | 4.00 | 3.49 | 4.00 | 4.00 | 0 (0.00) |
| I_03 Obeys the stay command immediately | 1.00 | 3.00 | 3.00 | 3.15 | 4.00 | 4.00 | 2 (0.03) |
| I_04 Seems to attend or listen closely to everything you say or do | 2.00 | 3.00 | 3.00 | 3.40 | 4.00 | 4.00 | 0 (0.00) |
| I_05 Slow to respond to correction or punishment thick-skinned | 0.00 | 3.00 | 3.00 | 2.96 | 4.00 | 4.00 | 6 (0.10) |
| I_06 Slow to learn new tricks or tasks | 0.00 | 3.00 | 3.00 | 3.16 | 4.00 | 4.00 | 0 (0.00) |
| I_07 Easily distracted by interesting sights sounds or smells | 0.00 | 2.00 | 2.00 | 2.16 | 3.00 | 4.00 | 0 (0.00) |
| I_08 Will fetch or attempt to fetch sticks balls or objects | 1.00 | 1.75 | 2.00 | 2.57 | 4.00 | 4.00 | 3 (0.05) |
| **Section 2. Aggression (0=No aggression, 4=Serious aggression)** | | | | | | | |
| I_09 When verbally corrected or punished scolded shouted at etc by you or a household member | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 6 (0.10) |
| I_10 When approached directly by an unfamiliar adult while being walked on a leash | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0 (0.00) |
| I_11 When approached directly by an unfamiliar child while being walked on a leash | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 1.00 | 0 (0.00) |
| I_12 Toward unfamiliar persons approaching the dog while she is in your car at the gas station for example | 0.00 | 0.00 | 0.00 | 0.06 | 0.00 | 2.00 | 1 (0.02) |
| I_13 When toys bones or other objects are taken away by a household member | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0 (0.00) |
| I_14 When bathed or groomed by a household member | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 1.00 | 0 (0.00) |
| I_15 When an unfamiliar person approaches you or another member of your family at home | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 2.00 | 0 (0.00) |

Continued on next page.

Table B.3 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| I_16 When unfamiliar persons approach you or another member of your family away from your home | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 1.00 | 0 (0.00) |
| I_17 When approached directly by a household member while she the dog is eating | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0 (0.00) |
| I_18 When mailmen or other delivery workers approach your home | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 2.00 | 1 (0.02) |
| I_19 When his or her food is taken away by a household member | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 1.00 | 3 (0.05) |
| I_20 When strangers walk past your home while your dog is outside or in the yard | 0.00 | 0.00 | 0.00 | 0.18 | 0.00 | 2.00 | 1 (0.02) |
| I_21 When an unfamiliar person tries to touch or pet the dog | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 1.00 | 0 (0.00) |
| I_22 When joggers cyclists rollerbladers or skateboarders pass your home | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 2.00 | 1 (0.02) |
| I_23 When approached directly by an unfamiliar male dog while being walked on a leash | 0.00 | 0.00 | 0.00 | 0.13 | 0.00 | 3.00 | 0 (0.00) |
| I_24 When approached directly by an unfamiliar female dog while being walked on a leash | 0.00 | 0.00 | 0.00 | 0.10 | 0.00 | 3.00 | 0 (0.00) |
| I_25 When stared at directly by a member of the household | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 3.00 | 0 (0.00) |
| I_26 Toward unfamiliar dogs visiting your home | 0.00 | 0.00 | 0.00 | 0.21 | 0.00 | 3.00 | 2 (0.03) |
| I_27 Toward cats squirrels or other small animals entering your yard | 0.00 | 0.00 | 0.00 | 0.38 | 1.00 | 4.00 | 2 (0.03) |
| I_28 Toward unfamiliar persons visiting your home | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 1.00 | 0 (0.00) |
| I_29 When barked growled or lunged at by another unfamiliar dog | 0.00 | 0.00 | 0.00 | 0.36 | 0.00 | 4.00 | 2 (0.03) |
| I_30 When stepped over by a member of the household | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0 (0.00) |
| I_31 When you or a household member retrieves food or objects stolen by the dog | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 1.00 | 2 (0.03) |

Continued on next page.

Table B.3 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| I_32 Towards another familiar dog in your household NA if no other dogs | 0.00 | 0.00 | 0.00 | 0.09 | 0.00 | 3.00 | 10 (0.16) |
| I_33 When approached at a favourite resting or sleeping place by another familiar household dog NA if no other dogs | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 1.00 | 11 (0.17) |
| I_34 When approached while eating by another familiar household dog NA if no other dogs | 0.00 | 0.00 | 0.00 | 0.10 | 0.00 | 2.00 | 12 (0.19) |
| I_35 When approached while playing with or chewing a favourite toy bone object etc by another familiar household dog NA if no other dogs | 0.00 | 0.00 | 0.00 | 0.09 | 0.00 | 3.00 | 10 (0.16) |
| **Section 3. Fear and Anxiety (0=No fear, 4=Extreme fear)** | | | | | | | |
| I_36 When approached directly by an unfamiliar adult while away from your home | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 1.00 | 0 (0.00) |
| I_37 When approached directly by an unfamiliar child while away from your home | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1 (0.02) |
| I_38 In response to sudden or loud noises eg vacuum cleaner car backfire road drills objects being dropped etc | 0.00 | 0.00 | 0.00 | 0.52 | 1.00 | 4.00 | 0 (0.00) |
| I_39 When unfamiliar persons visit your home | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 1.00 | 1 (0.02) |
| I_40 When an unfamiliar person tries to touch or pet the dog | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 1.00 | 0 (0.00) |
| I_41 In heavy traffic | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 2.00 | 0 (0.00) |
| I_42 In response to strange or unfamiliar objects on or near the sidewalk eg plastic trash bags leaves litter flags flapping etc | 0.00 | 0.00 | 0.00 | 0.18 | 0.00 | 2.00 | 1 (0.02) |
| I_43 When examined or treated by a veterinarian | 0.00 | 0.00 | 0.00 | 0.29 | 0.00 | 4.00 | 0 (0.00) |
| I_44 During thunderstorms firework displays or similar events | 0.00 | 0.00 | 0.00 | 0.22 | 0.00 | 2.00 | 3 (0.05) |
| I_45 When approached directly by an unfamiliar dog of the same or larger size | 0.00 | 0.00 | 0.00 | 0.30 | 0.00 | 4.00 | 0 (0.00) |

Continued on next page.

Table B.3 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| I_46 When approached directly by an unfamiliar dog of a smaller size | 0.00 | 0.00 | 0.00 | 0.24 | 0.00 | 4.00 | 0 (0.00) |
| I_47 When first exposed to unfamiliar situations eg first car trip first time in elevator first visit to veterinarian etc | 0.00 | 0.00 | 0.00 | 0.35 | 1.00 | 2.00 | 0 (0.00) |
| I_48 In response to wind or windblown objects | 0.00 | 0.00 | 0.00 | 0.24 | 0.00 | 4.00 | 0 (0.00) |
| I_49 When having nails clipped by a household member | 0.00 | 0.00 | 0.00 | 0.14 | 0.00 | 2.00 | 28 (0.44) |
| I_50 When groomed or bathed by a household member | 0.00 | 0.00 | 0.00 | 0.29 | 0.00 | 4.00 | 0 (0.00) |
| I_51 When having his or her feet towed by a member of the household | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 1.00 | 1 (0.02) |
| I_52 When unfamiliar dogs visit your home | 0.00 | 0.00 | 0.00 | 0.17 | 0.00 | 4.00 | 4 (0.06) |
| I_53 When barked growled or lunged at by an unfamiliar dog | 0.00 | 0.00 | 0.00 | 0.67 | 1.00 | 4.00 | 3 (0.05) |
| **Section 4. Separation-related behaviour (0=Never, 4=Always)** | | | | | | | |
| I_54 Shaking shivering or trembling | 0.00 | 0.00 | 0.00 | 0.10 | 0.00 | 2.00 | 1 (0.02) |
| I_55 Excessive salivation | 0.00 | 0.00 | 0.00 | 0.13 | 0.00 | 2.00 | 3 (0.05) |
| I_56 Restlessness agitation pacing | 0.00 | 0.00 | 0.00 | 0.35 | 0.00 | 3.00 | 1 (0.02) |
| I_57 Whining | 0.00 | 0.00 | 0.00 | 0.58 | 1.00 | 3.00 | 1 (0.02) |
| I_58 Howling | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 2.00 | 1 (0.02) |
| I_59 Chewing or scratching at doors floor windows curtains etc | 0.00 | 0.00 | 0.00 | 0.56 | 1.00 | 3.00 | 1 (0.02) |
| I_60 Loss of appetite | 0.00 | 0.00 | 0.00 | 0.07 | 0.00 | 2.00 | 2 (0.03) |
| I_61 Barking | 0.00 | 0.00 | 0.00 | 0.73 | 1.00 | 3.00 | 0 (0.00) |
| **Section 5. Excitability (0=Calm, 4=Extremely excitable)** | | | | | | | |
| I_62 When you or other members of the household come home after a brief absence | 0.00 | 1.00 | 2.00 | 1.83 | 3.00 | 4.00 | 0 (0.00) |
| I_63 When playing with you or other members of your household | 0.00 | 1.00 | 2.00 | 1.63 | 2.00 | 4.00 | 0 (0.00) |

Continued on next page.

Table B.3 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| I_64 When doorbell rings | 0.00 | 0.00 | 1.00 | 1.30 | 2.00 | 4.00 | 3 (0.05) |
| I_65 Just before being taken for a walk | 0.00 | 0.00 | 1.00 | 0.92 | 1.00 | 4.00 | 0 (0.00) |
| I_66 Just before being taken on a car trip | 0.00 | 0.00 | 0.00 | 0.65 | 1.00 | 4.00 | 0 (0.00) |
| I_67 When visitors arrive at your home | 0.00 | 1.00 | 1.00 | 1.54 | 2.00 | 4.00 | 0 (0.00) |
| I_68 Displays a strong attachment for one particular member of the household | 0.00 | 1.00 | 2.00 | 2.05 | 3.00 | 4.00 | 1 (0.02) |
| I_69 Tends to follow you or other members of household about the house from room to room | 0.00 | 2.00 | 3.00 | 2.52 | 3.00 | 4.00 | 0 (0.00) |
| I_70 Tends to sit close to or in contact with you or others when you are sitting down | 1.00 | 2.00 | 3.00 | 2.66 | 3.00 | 4.00 | 1 (0.02) |
| I_71 Tends to nudge nuzzle or paw you or others for attention when you are sitting down | 0.00 | 1.00 | 2.00 | 1.84 | 2.00 | 4.00 | 1 (0.02) |
| I_72 Becomes agitated whines jumps up tries to intervene when you or others show affection for another person | 0.00 | 0.00 | 0.00 | 0.68 | 1.00 | 3.00 | 1 (0.02) |
| I_73 Becomes agitated whines jumps up tries to intervene when you show affection for another dog or animal | 0.00 | 0.00 | 0.00 | 0.79 | 1.00 | 4.00 | 2 (0.03) |
| **Section 7. Miscellaneous (0=Never, 4=Always)** | | | | | | | |
| I_74 Chases or would chase cats given the opportunity | 0.00 | 0.00 | 1.00 | 1.48 | 2.00 | 4.00 | 7 (0.11) |
| I_75 Chases or would chase birds given the opportunity | 0.00 | 0.00 | 1.00 | 1.31 | 2.00 | 4.00 | 1 (0.02) |
| I_76 Chases or would chase squirrels rabbits and other small animals given the opportunity | 0.00 | 0.00 | 2.00 | 1.49 | 2.00 | 4.00 | 18 (0.29) |
| I_77 Escapes or would escape from home or yard given the chance | 0.00 | 0.00 | 0.00 | 0.65 | 1.00 | 4.00 | 6 (0.10) |
| I_78 Rolls in animal droppings or other smelly substances | 0.00 | 0.00 | 1.00 | 1.13 | 2.00 | 4.00 | 2 (0.03) |
| I_79 Eats own or other animals droppings or faeces | 0.00 | 0.00 | 0.00 | 0.87 | 2.00 | 4.00 | 0 (0.00) |

Continued on next page.

Table B.3 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| I_80 Chews inappropriate objects | 0.00 | 0.00 | 1.00 | 1.11 | 2.00 | 4.00 | 2 (0.03) |
| I_81 Mounts objects furniture or people | 0.00 | 0.00 | 0.00 | 0.27 | 0.00 | 2.00 | 1 (0.02) |
| I_82 Begs persistently for food when people are eating | 0.00 | 0.00 | 0.00 | 0.61 | 1.00 | 4.00 | 1 (0.02) |
| I_83 Steals food | 0.00 | 0.00 | 0.00 | 0.77 | 1.00 | 4.00 | 1 (0.02) |
| I_84 Nervous or frightened on stairs | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 | 4.00 | 1 (0.02) |
| I_85 Pulls excessively hard when on the leash | 0.00 | 0.00 | 1.00 | 0.94 | 2.00 | 3.00 | 1 (0.02) |
| I_86 Urinates against objects furnishings in your home | 0.00 | 0.00 | 0.00 | 0.06 | 0.00 | 4.00 | 1 (0.02) |
| I_87 Urinates when approached petted handled or picked up | 0.00 | 0.00 | 0.00 | 0.06 | 0.00 | 4.00 | 1 (0.02) |
| I_88 Urinates when left alone at night or during the daytime | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 4.00 | 1 (0.02) |
| I_89 Defecates when left alone at night or during the daytime | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 4.00 | 1 (0.02) |
| I_90 Hyperactive restless has trouble settling down | 0.00 | 0.00 | 0.00 | 0.32 | 0.00 | 3.00 | 1 (0.02) |
| I_91 Playful puppyish boisterous | 0.00 | 1.00 | 2.00 | 1.68 | 2.00 | 4.00 | 0 (0.00) |
| I_92 Active energetic always on the go | 0.00 | 1.00 | 2.00 | 1.75 | 3.00 | 4.00 | 0 (0.00) |
| I_93 Stares intently at nothing visible | 0.00 | 0.00 | 0.00 | 0.23 | 0.00 | 4.00 | 7 (0.11) |
| I_94 Snaps at invisible flies | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 4.00 | 4 (0.06) |
| I_95 Chases own tail or hind end | 0.00 | 0.00 | 1.00 | 0.92 | 2.00 | 3.00 | 2 (0.03) |
| I_96 Chases or follows shadows light spots etc | 0.00 | 0.00 | 0.00 | 0.45 | 0.00 | 4.00 | 5 (0.08) |
| I_97 Barks persistently when alarmed or excited | 0.00 | 0.00 | 0.00 | 0.49 | 1.00 | 4.00 | 0 (0.00) |
| I_98 Licks him or herself excessively | 0.00 | 0.00 | 0.00 | 0.34 | 1.00 | 2.00 | 2 (0.03) |
| I_99 Licks people or objects excessively | 0.00 | 0.00 | 0.00 | 0.47 | 1.00 | 3.00 | 3 (0.05) |
| I_100 Displays other bizarre strange or repetitive behaviours | 0.00 | 0.00 | 0.00 | 0.18 | 0.00 | 4.00 | 7 (0.11) |
| **Factors** | | | | | | | |
| F_01 Trainability | 2.12 | 2.75 | 3.00 | 3.00 | 3.25 | 3.88 | 0 (0.00) |
| F_02 Stranger directed aggression | 0.00 | 0.00 | 0.00 | 0.06 | 0.10 | 0.50 | 0 (0.00) |
| F_03 Owner directed aggression | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.43 | 0 (0.00) |

Continued on next page.

Table B.3 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| F_04 Dog rivalry | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 1.50 | 11 (0.17) |
| F_05 Stranger directed fear | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.75 | 0 (0.00) |
| F_06 Nonsocial fear | 0.00 | 0.00 | 0.17 | 0.29 | 0.50 | 1.60 | 0 (0.00) |
| F_07 Dog directed aggression | 0.00 | 0.00 | 0.00 | 0.20 | 0.00 | 3.25 | 0 (0.00) |
| F_08 Dog directed fear | 0.00 | 0.00 | 0.00 | 0.34 | 0.29 | 3.75 | 0 (0.00) |
| F_09 Touch sensitivity | 0.00 | 0.00 | 0.00 | 0.16 | 0.25 | 1.25 | 1 (0.02) |
| F_10 Separation related problems | 0.00 | 0.00 | 0.19 | 0.34 | 0.62 | 1.50 | 1 (0.02) |
| F_11 Excitability | 0.00 | 0.67 | 1.00 | 1.31 | 1.92 | 3.83 | 0 (0.00) |
| F_12 Attachment attention seeking | 0.33 | 1.33 | 1.83 | 1.76 | 2.17 | 3.50 | 2 (0.03) |
| F_13 Chasing | 0.00 | 0.29 | 1.00 | 1.13 | 1.50 | 4.00 | 4 (0.06) |
| F_14 Energy | 0.00 | 1.00 | 1.50 | 1.71 | 2.25 | 4.00 | 0 (0.00) |

### B.1.2 Univariate Analysis

Table B.4: Univariate logistic regression model estimates, p-value, Odds Ratio (OR), and Confidence Interval (CI).

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| * I_01 When off the leash returns immediately when called | -1.12 | 0.01 | 0.33 | 0.12 | 0.76 |
| I_02 Obeys the sit command immediately | -0.39 | 0.45 | 0.68 | 0.24 | 1.95 |
| I_03 Obeys the stay command immediately | 0.11 | 0.81 | 1.11 | 0.46 | 2.91 |
| I_04 Seems to attend or listen closely to everything you say or do | -0.12 | 0.83 | 0.89 | 0.29 | 2.81 |
| [b] I_05 Slow to respond to correction or punishment thick-skinned | -0.08 | 0.83 | 0.93 | 0.48 | 1.98 |
| I_06 Slow to learn new tricks or tasks | 0.17 | 0.66 | 1.18 | 0.60 | 2.79 |
| I_07 Easily distracted by interesting sights sounds or smells | -0.78 | 0.06 | 0.46 | 0.19 | 1.02 |
| I_08 Will fetch or attempt to fetch sticks balls or objects | -0.06 | 0.85 | 0.94 | 0.52 | 1.69 |
| [a,b] I_09 When verbally corrected or punished scolded shouted at etc by you or a household member | 0.35 | -4.44 | 1.42 | | |
| [a] I_10 When approached directly by an unfamiliar adult while being walked on a leash | 0.33 | -4.68 | 1.39 | | |
| I_11 When approached directly by an unfamiliar child while being walked on a leash | 18.21 | 0.99 | | | |
| I_12 Toward unfamiliar persons approaching the dog while she is in your car at the gas station for example | 1.09 | 0.22 | 2.97 | 0.47 | 22.66 |
| [a] I_13 When toys bones or other objects are taken away by a household member | 0.33 | -4.68 | 1.39 | | |
| I_14 When bathed or groomed by a household member | -15.03 | 1.00 | | | |

Continued on next page.

Table B.4 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| I_15 When an unfamiliar person approaches you or another member of your family at home | 9.11 | 0.99 | | | |
| I_16 When unfamiliar persons approach you or another member of your family away from your home | 18.21 | 0.99 | | | |
| [a] I_17 When approached directly by a household member while she the dog is eating | 0.33 | -4.68 | 1.39 | | |
| I_18 When mailmen or other delivery workers approach your home | 0.38 | 0.54 | 1.47 | 0.35 | 4.80 |
| I_19 When his or her food is taken away by a household member | -14.98 | 1.00 | | | |
| I_20 When strangers walk past your home while your dog is outside or in the yard | 0.03 | 0.97 | 1.03 | 0.16 | 4.12 |
| I_21 When an unfamiliar person tries to touch or pet the dog | 18.21 | 0.99 | | | |
| I_22 When joggers cyclists rollerbladers or skateboarders pass your home | 0.86 | 0.29 | 2.37 | 0.38 | 13.30 |
| I_23 When approached directly by an unfamiliar male dog while being walked on a leash | 1.01 | 0.12 | 2.75 | 0.82 | 12.94 |
| I_24 When approached directly by an unfamiliar female dog while being walked on a leash | 0.83 | 0.20 | 2.30 | 0.62 | 11.63 |
| I_25 When stared at directly by a member of the household | 6.07 | 0.99 | 433.36 | 0.00 | |
| I_26 Toward unfamiliar dogs visiting your home | 1.31 | 0.01 | 3.70 | 1.51 | 12.95 |
| * I_27 Toward cats squirrels or other small animals entering your yard | 1.00 | 0.04 | 2.73 | 1.15 | 8.14 |
| I_28 Toward unfamiliar persons visiting your home | 1.63 | 0.26 | 5.10 | 0.19 | 136.44 |

Continued on next page.

Table B.4 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| I_29 When barked growled or lunged at by another unfamiliar dog | 1.07 | 0.01 | 2.92 | 1.43 | 8.11 |
| [a] I_30 When stepped over by a member of the household | 0.33 | -4.68 | 1.39 | | |
| I_31 When you or a household member retrieves food or objects stolen by the dog | -14.98 | 0.99 | | | |
| [b] I_32 Towards another familiar dog in your household NA if no other dogs | 1.93 | 0.13 | 6.90 | 1.21 | 172.19 |
| [b] I_33 When approached at a favourite resting or sleeping place by another familiar household dog NA if no other dogs | -15.05 | 0.99 | | | |
| [b] I_34 When approached while eating by another familiar household dog NA if no other dogs | -16.30 | 1.00 | | | |
| [b] I_35 When approached while playing with or chewing a favourite toy bone object etc by another familiar household dog NA if no other dogs | 0.93 | 0.20 | 2.54 | 0.67 | 18.11 |
| I_36 When approached directly by an unfamiliar adult while away from your home | 0.18 | 0.88 | 1.20 | 0.06 | 9.26 |
| [a] I_37 When approached directly by an unfamiliar child while away from your home | 0.33 | -4.61 | 1.39 | | |
| I_38 In response to sudden or loud noises eg vacuum cleaner car backfire road drills objects being dropped etc | 1.03 | 0.03 | 2.80 | 1.25 | 8.20 |
| I_39 When unfamiliar persons visit your home | 1.61 | 0.27 | 5.00 | 0.19 | 133.78 |
| I_40 When an unfamiliar person tries to touch or pet the dog | 18.32 | 0.99 | | | |
| I_41 In heavy traffic | 1.12 | 0.06 | 3.08 | 0.94 | 10.64 |

Continued on next page.

Table B.4 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| I_42 In response to strange or unfamiliar objects on or near the sidewalk eg plastic trash bags leaves litter flags flapping etc | 0.68 | 0.33 | 1.97 | 0.43 | 7.59 |
| I_43 When examined/treated by a veterinarian | 0.91 | 0.05 | 2.47 | 1.06 | 7.04 |
| I_44 During thunderstorms firework displays or similar events | 0.23 | 0.69 | 1.26 | 0.32 | 3.76 |
| I_45 When approached directly by an unfamiliar dog of the same or larger size | 0.83 | 0.03 | 2.30 | 1.16 | 5.70 |
| I_46 When approached directly by an unfamiliar dog of a smaller size | 0.70 | 0.08 | 2.01 | 0.93 | 5.01 |
| I_47 When first exposed to unfamiliar situations eg first car trip first time in elevator first visit to veterinarian etc | 0.92 | 0.08 | 2.52 | 0.88 | 7.33 |
| I_48 In response to wind or windblown objects | 1.06 | 0.08 | 2.88 | 1.07 | 11.19 |
| [b] I_49 When having nails clipped by a household member | 1.29 | 0.10 | 3.62 | 0.63 | 19.43 |
| I_50 When groomed or bathed by a household member | 0.47 | 0.22 | 1.60 | 0.71 | 3.52 |
| I_51 When having his or her feet towed by a member of the household | -14.96 | 0.99 | | | |
| [b] I_52 When unfamiliar dogs visit your home | 1.39 | 0.06 | 4.03 | 1.28 | 22.55 |
| I_53 When barked growled or lunged at by an unfamiliar dog | 0.70 | 0.02 | 2.02 | 1.13 | 4.00 |
| I_54 Shaking shivering or trembling | 1.51 | 0.05 | 4.53 | 1.14 | 30.34 |
| I_55 Excessive salivation | 1.14 | 0.12 | 3.12 | 0.67 | 14.31 |
| I_56 Restlessness agitation pacing | 0.81 | 0.03 | 2.25 | 1.05 | 4.91 |
| I_57 Whining | 0.79 | 0.03 | 2.20 | 1.07 | 4.74 |
| I_58 Howling | 0.84 | 0.10 | 2.33 | 0.82 | 6.52 |

Continued on next page.

Table B.4 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| I_59 Chewing or scratching at doors floor windows curtains etc | 0.32 | 0.32 | 1.38 | 0.70 | 2.58 |
| I_60 Loss of appetite | 1.16 | 0.19 | 3.20 | 0.50 | 24.82 |
| I_61 Barking | 0.66 | 0.03 | 1.93 | 1.07 | 3.60 |
| I_62 When you or other members of the household come home after a brief absence | 1.04 | 0.00 | 2.84 | 1.49 | 6.20 |
| I_63 When playing with you or other members of your household | 0.88 | 0.01 | 2.40 | 1.27 | 5.11 |
| I_64 When doorbell rings | 0.75 | 0.02 | 2.12 | 1.18 | 4.11 |
| I_65 Just before being taken for a walk | 0.73 | 0.01 | 2.07 | 1.19 | 3.85 |
| I_66 Just before being taken on a car trip | 0.64 | 0.03 | 1.90 | 1.06 | 3.53 |
| I_67 When visitors arrive at your home | 0.46 | 0.09 | 1.58 | 0.92 | 2.76 |
| I_68 Displays a strong attachment for one particular member of the household | 0.40 | 0.16 | 1.50 | 0.88 | 2.79 |
| I_69 Tends to follow you or other members of household about the house from room to room | 0.54 | 0.16 | 1.72 | 0.85 | 3.87 |
| I_70 Tends to sit close to or in contact with you or others when you are sitting down | 0.24 | 0.52 | 1.27 | 0.61 | 2.80 |
| I_71 Tends to nudge nuzzle or paw you or others for attention when you are sitting down | 0.41 | 0.17 | 1.51 | 0.85 | 2.83 |
| I_72 Becomes agitated whines jumps up tries to intervene when you or others show affection for another person | 0.46 | 0.15 | 1.58 | 0.83 | 2.98 |
| I_73 Becomes agitated whines jumps up tries to intervene when you show affection for another dog or animal | 0.34 | 0.27 | 1.41 | 0.75 | 2.60 |
| [b] I_74 Chases or would chase cats given the opportunity | 0.15 | 0.57 | 1.16 | 0.69 | 1.92 |

Continued on next page.

Table B.4 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| * I_75 Chases or would chase birds given the opportunity | 0.61 | 0.03 | 1.84 | 1.09 | 3.24 |
| [b] I_76 Chases or would chase squirrels rabbits and other small animals given the opportunity | 0.32 | 0.23 | 1.38 | 0.82 | 2.40 |
| * [b] I_77 Escapes or would escape from home or yard given the chance | 1.05 | 0.01 | 2.86 | 1.44 | 6.84 |
| I_78 Rolls in animal droppings or other smelly substances | -0.04 | 0.90 | 0.96 | 0.52 | 1.69 |
| I_79 Eats own or other animals droppings or faeces | -0.36 | 0.31 | 0.70 | 0.30 | 1.29 |
| I_80 Chews inappropriate objects | 0.37 | 0.24 | 1.44 | 0.78 | 2.72 |
| I_81 Mounts objects furniture or people | 0.30 | 0.57 | 1.35 | 0.42 | 3.64 |
| * I_82 Begs persistently for food when people are eating | 0.68 | 0.03 | 1.98 | 1.10 | 3.80 |
| I_83 Steals food | 0.04 | 0.89 | 1.04 | 0.54 | 1.79 |
| I_84 Nervous or frightened on stairs | 0.07 | 0.90 | 1.07 | 0.23 | 2.69 |
| * I_85 Pulls excessively hard when on the leash | 0.82 | 0.04 | 2.27 | 1.07 | 5.31 |
| I_86 Urinates against objects furnishings in your home | -3.76 | 0.99 | | | |
| I_87 Urinates when approached petted handled or picked up | -3.76 | 0.99 | | | |
| I_88 Urinates when left alone at night or during the daytime | -14.88 | 0.99 | | | |
| I_89 Defecates when left alone at night or during the daytime | -14.88 | 0.99 | | | |
| * I_90 Hyperactive restless has trouble settling down | 0.87 | 0.04 | 2.39 | 1.03 | 5.86 |
| I_91 Playful puppyish boisterous | 0.66 | 0.07 | 1.93 | 0.98 | 4.10 |
| I_92 Active energetic always on the go | 0.57 | 0.08 | 1.77 | 0.95 | 3.51 |
| * [b] I_93 Stares intently at nothing visible | 1.03 | 0.05 | 2.80 | 1.14 | 9.28 |
| [b] I_94 Snaps at invisible flies | 0.41 | 0.31 | 1.51 | 0.60 | 3.57 |
| I_95 Chases own tail or hind end | 0.38 | 0.29 | 1.46 | 0.72 | 3.04 |

Continued on next page.

Table B.4 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| [b] I_96 Chases or follows shadows light spots etc | 0.43 | 0.12 | 1.54 | 0.87 | 2.71 |
| I_97 Barks persistently when alarmed or excited | 0.26 | 0.50 | 1.30 | 0.55 | 2.75 |
| I_98 Licks him or herself excessively | 0.80 | 0.13 | 2.24 | 0.75 | 6.59 |
| * I_99 Licks people or objects excessively | 0.84 | 0.03 | 2.32 | 1.08 | 5.22 |
| [b] I_100 Displays other bizarre strange or repetitive behaviours | 0.08 | 0.85 | 1.09 | 0.28 | 2.31 |
| F_01 Trainability | -1.03 | 0.21 | 0.36 | 0.07 | 1.80 |
| * F_02 Stranger directed aggression | 5.74 | 0.03 | 311.63 | 1.94 | 78300.96 |
| F_03 Owner directed aggression | 4.85 | 0.23 | 127.29 | 0.02 | 1356011.42 |
| [b] F_04 Dog rivalry | 1.25 | 0.23 | 3.49 | 0.37 | 33.64 |
| F_05 Stranger directed fear | 3.35 | 0.15 | 28.60 | 0.27 | 5492.23 |
| * F_06 Nonsocial fear | 2.51 | 0.01 | 12.32 | 2.18 | 95.69 |
| * F_07 Dog directed aggression | 1.58 | 0.02 | 4.84 | 1.51 | 21.91 |
| * F_08 Dog directed fear | 0.96 | 0.03 | 2.61 | 1.19 | 7.54 |
| F_09 Touch sensitivity | 0.85 | 0.43 | 2.35 | 0.21 | 18.47 |
| * F_10 Separation related problems | 2.85 | 0.00 | 17.24 | 3.08 | 135.53 |
| * F_11 Excitability | 1.25 | 0.00 | 3.51 | 1.60 | 9.03 |
| F_12 Attachment attention seeking | 1.00 | 0.08 | 2.72 | 0.96 | 9.27 |
| [b] F_13 Chasing | 0.62 | 0.07 | 1.87 | 0.96 | 3.82 |
| F_14 Energy | 0.75 | 0.05 | 2.11 | 1.03 | 4.73 |

95% Confidence Interval (CI): lower 2.5% and upper 97.5%
[a] No variance, constant values
[b] More than 5% missing data (4 data points)
* Statistically significant with $p < 0.05$. Variables that satisfied criteria 1 and 2.

### B.1.3 Visualisation

The C-BARQ items and factors selected for learning are displayed in Figures B.1 to B.12.

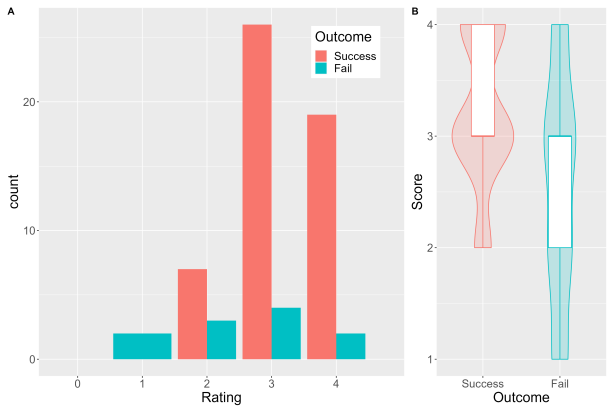



Figure B.1: Bar chart and box plot of the distribution of the C-BARQ item "I_01 When off the leash returns immediately when called" in relation to the working outcome.

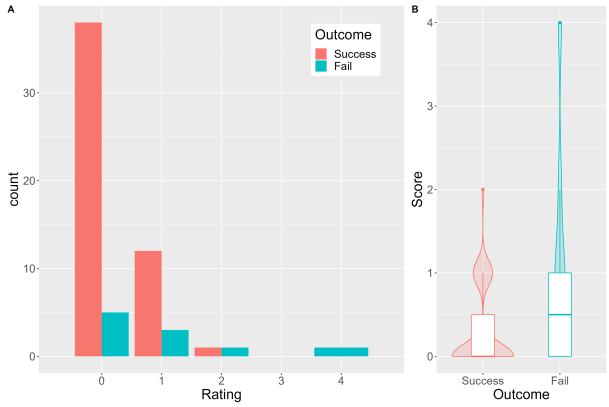



Figure B.2: Bar chart and box plot of the distribution of the C-BARQ item "I_27 Toward cats squirrels or other small animals entering your yard" in relation to the working outcome.

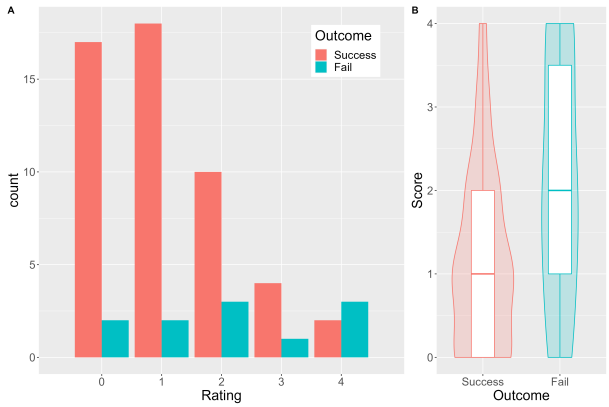



Figure B.3: Bar chart and box plot of the distribution of the C-BARQ item "I_75 Chases or would chase birds given the opportunity" in relation to the working outcome.

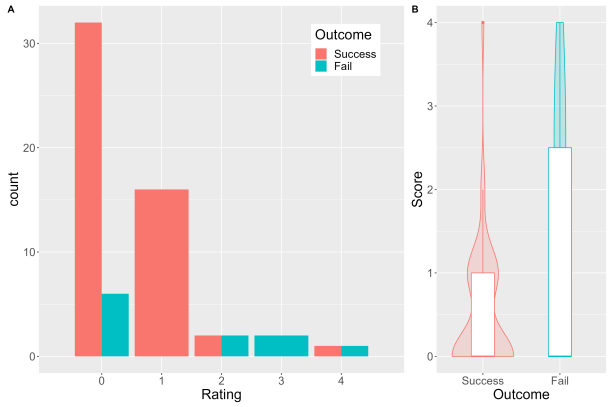



Figure B.4: Bar chart and box plot of the distribution of the C-BARQ item "I_82 Begs persistently for food when people are eating" in relation to the working outcome.

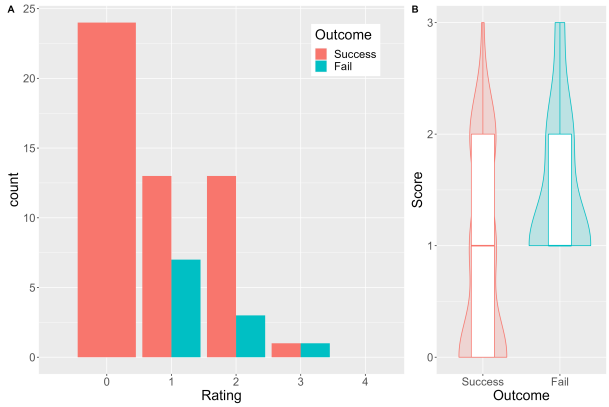



Figure B.5: Bar chart and box plot of the distribution of the C-BARQ item "I_85 Pulls excessively hard when on the leash" in relation to the working outcome.

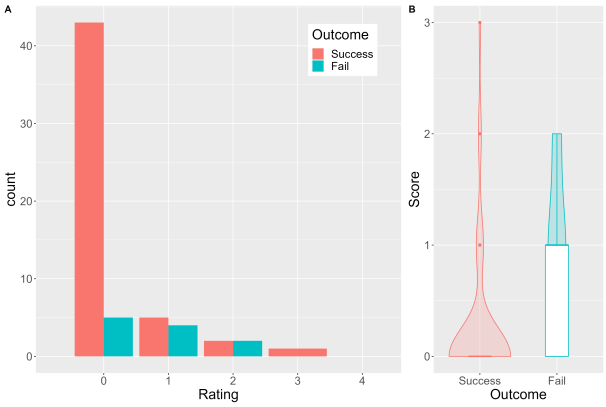



Figure B.6: Bar chart and box plot of the distribution of the C-BARQ item "I_90 Hyperactive restless has trouble settling down" in relation to the working outcome.

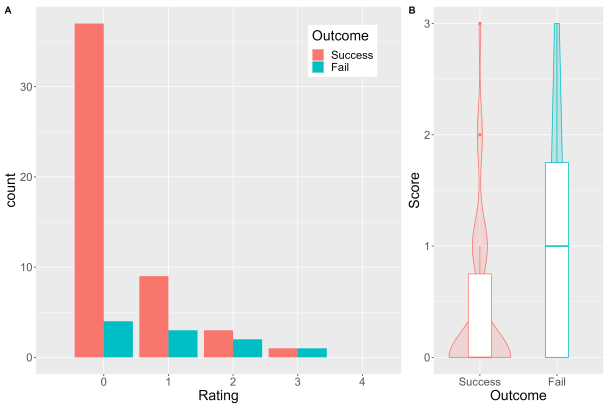



Figure B.7: Bar chart and box plot of the distribution of the C-BARQ item "I_99 Licks people or objects excessively.

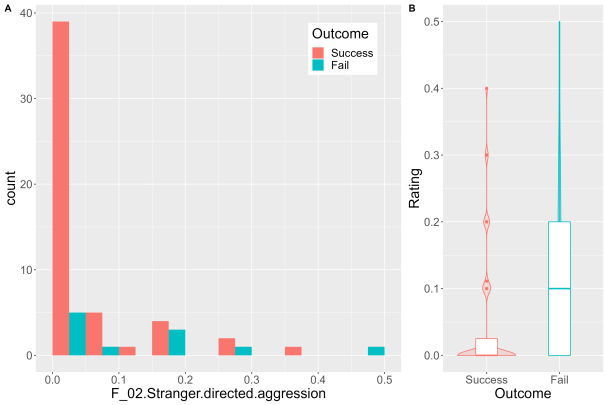



Figure B.8: Bar chart and box plot of the distribution of the C-BARQ factor "F_02 Stranger directed aggression" in relation to the working outcome.

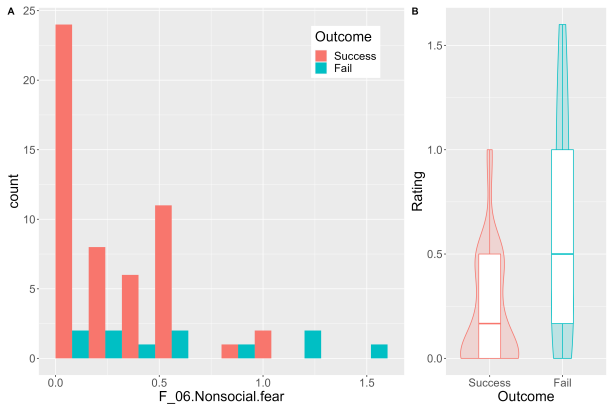



Figure B.9: Bar chart and box plot of the distribution of the C-BARQ factor "F_06 Nonsocial fear" in relation to the working outcome.

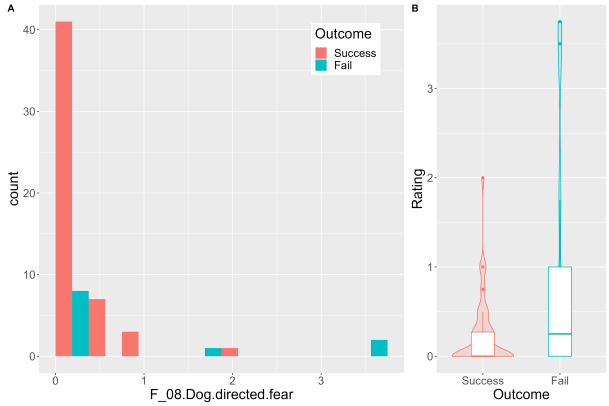



Figure B.10: Bar chart and box plot of the distribution of the C-BARQ factor "F_08 Dog directed fear."

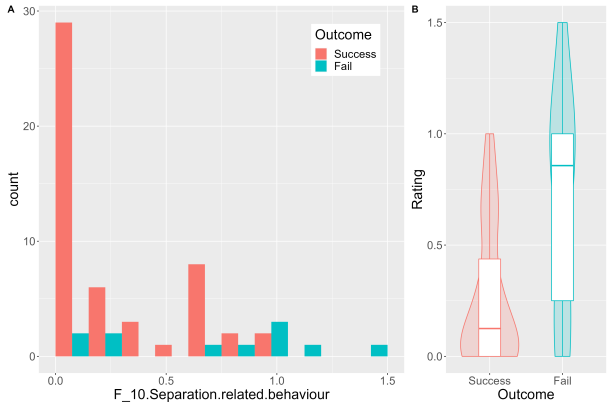



Figure B.11: Bar chart and box plot of the distribution of the C-BARQ factor "F_10 Separation related behaviour."

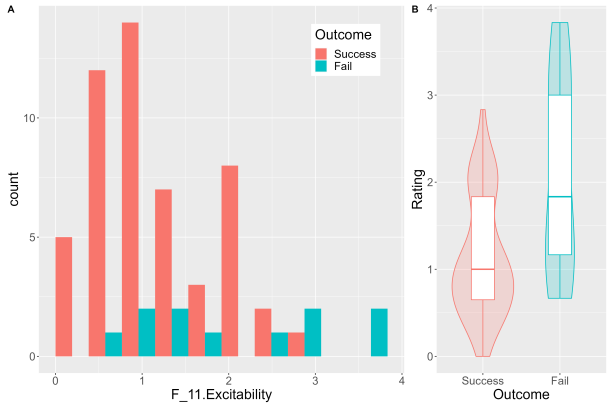



Figure B.12: Bar chart and box plot of the distribution of the C-BARQ factor "F_11 Excitability."

## B.2 Monash Canine Personality Questionnaire - Revised (MCPQ-R)

This questionnaire was developed, standardised, and validated by Ley et al. 2008; Ley et al. 2009a; Ley et al. 2009b. Respondents rate their dogs in a 6-point scale from 1 (really does not describe my dog) to 6 (really describes my dog).

Table B.5: MCPQ-R rating sub-scales and items.

| Extraversion | Motivation | Training | Amicability | Neuroticism |
|---|---|---|---|---|
| Active | Assertive | Attentive | Easy going | Fearful |
| Energetic | Determined | Biddable | Friendly | Nervous |
| Excitable | Independent | Intelligent | Non-aggressive | Submissive |
| Hyperactive | Persevering | Obedient | Relaxed | Timid |
| Lively | Tenacious | Reliable | Sociable | |
| Restless | | Trainable | | |

### B.2.1 Descriptive Statistics

Table B.6: MCPQ-R rating descriptive statistics including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing values for all items and factors.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max |
|---|---|---|---|---|---|---|
| Extraversion Active | 1.00 | 3.00 | 4.00 | 3.81 | 5.00 | 6.00 |
| Extraversion Energetic | 1.00 | 3.00 | 4.00 | 3.76 | 5.00 | 6.00 |
| Extraversion Excitable | 1.00 | 3.00 | 4.00 | 3.75 | 5.00 | 6.00 |
| Extraversion Hyperactive | 1.00 | 2.00 | 3.00 | 2.72 | 3.00 | 6.00 |
| Extraversion Lively | 1.00 | 3.00 | 4.00 | 3.62 | 4.00 | 6.00 |
| Extraversion Restless | 1.00 | 2.00 | 3.00 | 3.07 | 4.00 | 6.00 |
| Motivation Assertive | 1.00 | 3.00 | 3.00 | 3.24 | 4.00 | 6.00 |
| Motivation Determined | 1.00 | 3.00 | 4.00 | 3.55 | 4.00 | 5.00 |
| Motivation Independent | 1.00 | 2.00 | 3.00 | 3.25 | 4.00 | 5.00 |
| Motivation Persevering | 1.00 | 3.00 | 3.00 | 3.40 | 4.00 | 5.00 |
| Motivation Tenacious | 1.00 | 3.00 | 3.00 | 3.22 | 4.00 | 6.00 |
| Training Attentive | 1.00 | 3.00 | 4.00 | 3.91 | 5.00 | 6.00 |
| Training Biddable | 2.00 | 3.00 | 4.00 | 3.85 | 5.00 | 6.00 |
| Training Intelligent | 2.00 | 4.00 | 5.00 | 4.49 | 5.00 | 6.00 |
| Training Obedient | 2.00 | 3.00 | 4.00 | 4.07 | 5.00 | 6.00 |
| Training Reliable | 2.00 | 3.00 | 4.00 | 3.80 | 5.00 | 6.00 |
| Training Trainable | 1.00 | 4.00 | 5.00 | 4.37 | 5.00 | 6.00 |
| Amicability Easy going | 1.00 | 3.00 | 4.00 | 3.69 | 5.00 | 6.00 |
| Amicability Friendly | 1.00 | 4.00 | 5.00 | 4.93 | 6.00 | 6.00 |
| Amicability Non aggressive | 1.00 | 5.00 | 6.00 | 5.11 | 6.00 | 6.00 |
| Amicability Relaxed | 1.00 | 3.00 | 4.00 | 3.65 | 5.00 | 6.00 |
| Amicability Sociable | 1.00 | 4.00 | 5.00 | 4.67 | 5.00 | 6.00 |
| Neuroticism Fearful | 1.00 | 2.00 | 2.00 | 2.37 | 3.00 | 6.00 |
| Neuroticism Nervous | 1.00 | 1.00 | 2.00 | 2.38 | 3.00 | 6.00 |
| Neuroticism Submissive | 1.00 | 2.00 | 2.00 | 2.62 | 3.00 | 5.00 |
| Neuroticism Timid | 1.00 | 2.00 | 2.00 | 2.51 | 3.00 | 6.00 |

### B.2.2 Univariate Analysis

Table B.7: Univariate logistic regression model estimates, p-value, Odds Ratio (OR), and Confidence Interval (CI).

| | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| Extraversion Active | -0.22 | 0.38 | 0.80 | 0.48 | 1.31 |
| Extraversion Energetic | -0.14 | 0.54 | 0.87 | 0.55 | 1.35 |
| Extraversion Excitable | 0.17 | 0.45 | 1.18 | 0.77 | 1.82 |
| Extraversion Hyperactive | 0.37 | 0.07 | 1.45 | 0.97 | 2.18 |
| Extraversion Lively | -0.01 | 0.98 | 0.99 | 0.64 | 1.52 |
| Extraversion Restless | 0.18 | 0.35 | 1.19 | 0.82 | 1.74 |
| Motivation Assertive | -0.10 | 0.72 | 0.90 | 0.50 | 1.58 |
| Motivation Determined | -0.48 | 0.10 | 0.62 | 0.34 | 1.08 |
| Motivation Independent | -0.32 | 0.20 | 0.73 | 0.44 | 1.17 |
| * Motivation Persevering | -0.59 | 0.05 | 0.55 | 0.30 | 0.97 |
| Motivation Tenacious | -0.19 | 0.45 | 0.82 | 0.49 | 1.35 |
| * Training Attentive | -0.56 | 0.03 | 0.57 | 0.34 | 0.93 |
| Training Biddable | -0.49 | 0.07 | 0.62 | 0.36 | 1.02 |
| Training Intelligent | -0.39 | 0.13 | 0.68 | 0.41 | 1.12 |
| * Training Obedient | -0.63 | 0.02 | 0.53 | 0.31 | 0.89 |
| * Training Reliable | -1.09 | 0.00 | 0.34 | 0.17 | 0.60 |
| * Training Trainable | -1.09 | 0.00 | 0.34 | 0.18 | 0.57 |
| * Amicability Easy going | -0.84 | 0.00 | 0.43 | 0.24 | 0.70 |
| * Amicability Friendly | -0.76 | 0.00 | 0.47 | 0.27 | 0.76 |
| Amicability Non aggressive | -0.15 | 0.43 | 0.86 | 0.61 | 1.27 |
| * Amicability Relaxed | -0.73 | 0.00 | 0.48 | 0.29 | 0.74 |
| * Amicability Sociable | -0.64 | 0.01 | 0.53 | 0.31 | 0.84 |
| * Neuroticism Fearful | 0.74 | 0.00 | 2.10 | 1.37 | 3.39 |
| * Neuroticism Nervous | 0.65 | 0.00 | 1.91 | 1.28 | 2.98 |
| * Neuroticism Submissive | 0.64 | 0.03 | 1.90 | 1.09 | 3.44 |
| * Neuroticism Timid | 0.71 | 0.01 | 2.03 | 1.25 | 3.48 |

95% Confidence Interval (CI): lower 2.5% and upper 97.5%
* Statistically significant with p < 0.05. Variables that satisfied criteria 1 and 2.

### B.2.3 Visualisation

The MCPQ-R items selected for learning are displayed in Figures B.13 to B.24.

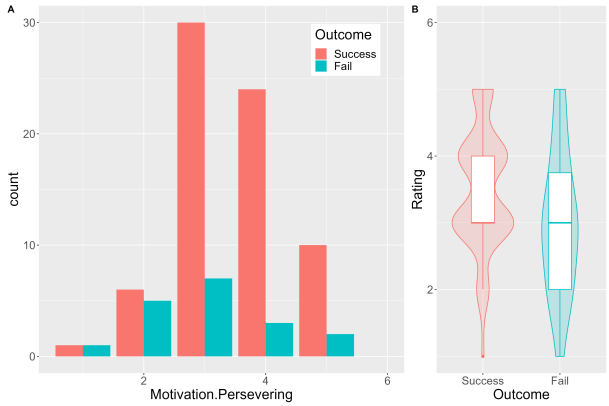



Figure B.13: Bar chart and box plot of the distribution of the MCPQ-R item "Motivation Persevering" in relation to training outcomes.

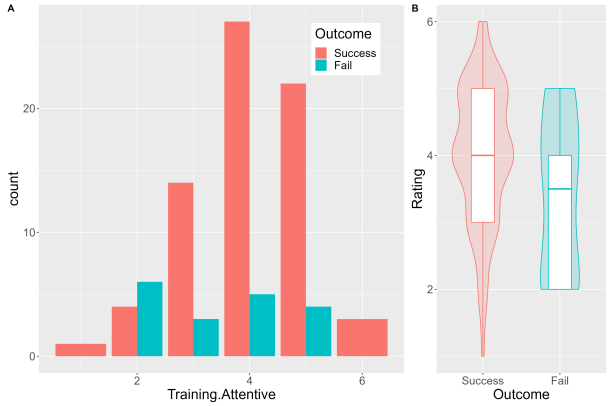



Figure B.14: Bar chart and box plot of the distribution of the MCPQ-R item "Training Attentive" in relation to training outcomes.

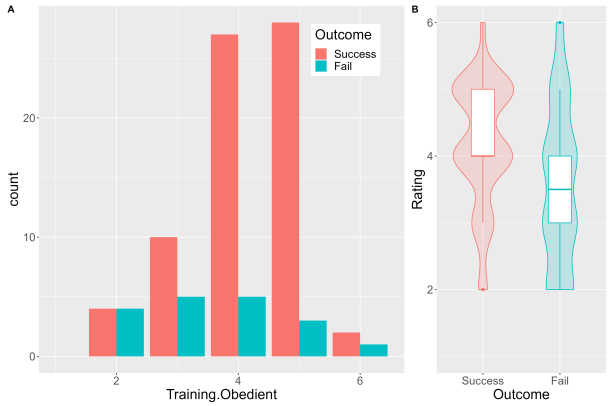



Figure B.15: Bar chart and box plot of the distribution of the MCPQ-R item "Training Obedient" in relation to training outcomes.

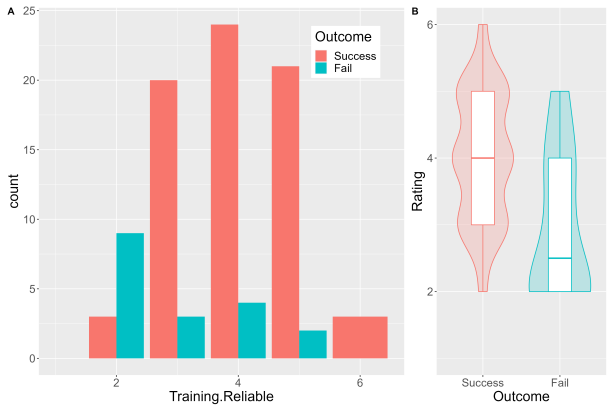



Figure B.16: Bar chart and box plot of the distribution of the MCPQ-R item "Training Reliable" in relation to training outcomes.

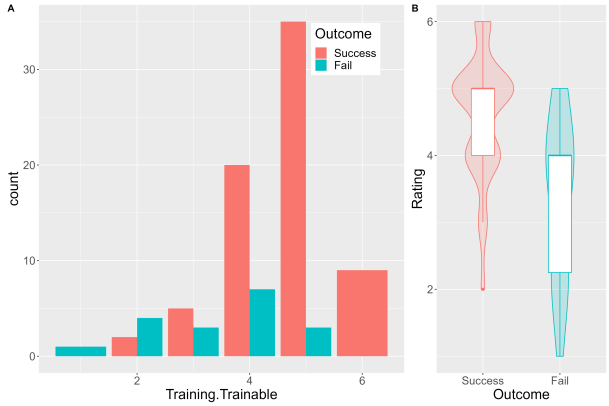



Figure B.17: Bar chart and box plot of the distribution of the MCPQ-R item "Training Trainable" in relation to training outcomes.

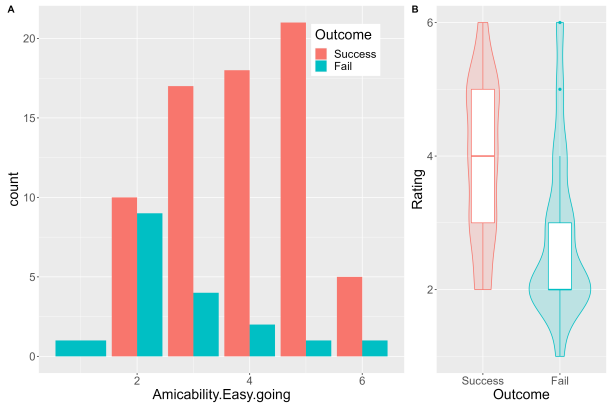



Figure B.18: Bar chart and box plot of the distribution of the MCPQ-R item "Amicability Easy-going" in relation to training outcomes.

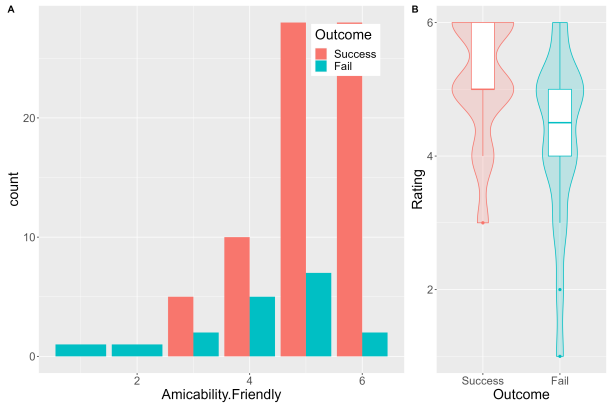



Figure B.19: Bar chart and box plot of the distribution of the MCPQ-R item "Amicability Friendly" in relation to training outcomes.

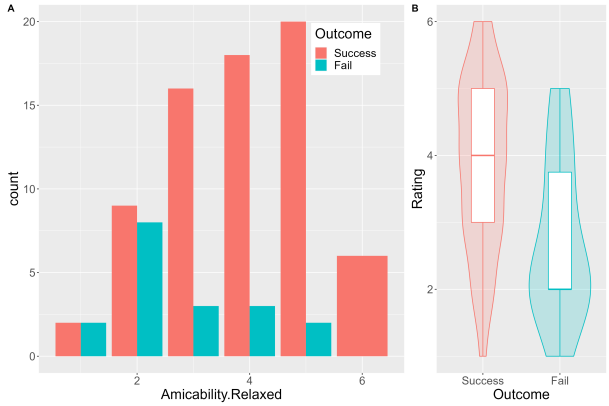



Figure B.20: Bar chart and box plot of the distribution of the MCPQ-R item "Amicability Relaxed" in relation to training outcomes.

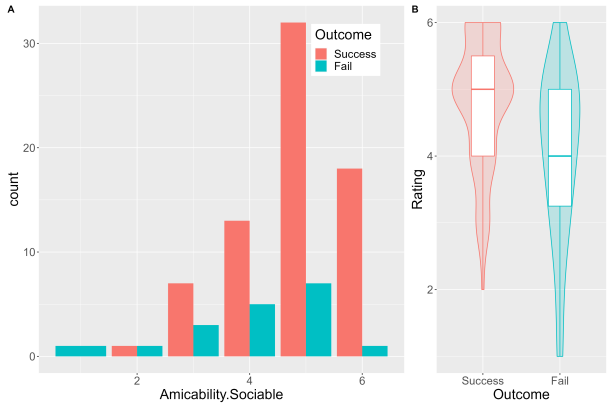



Figure B.21: Bar chart and box plot of the distribution of the MCPQ-R item "Amicability Sociable" in relation to training outcomes.

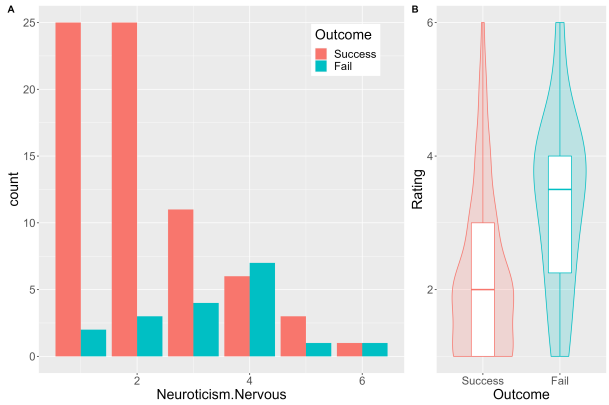



Figure B.22: Bar chart and box plot of the distribution of the MCPQ-R item "Neuroticism Nervous" in relation to training outcomes.

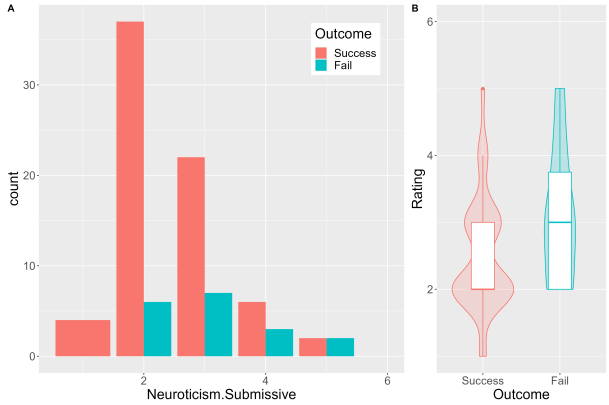



Figure B.23: Bar chart and box plot of the distribution of the MCPQ-R item "Neuroticism Submissive" in relation to training outcomes.

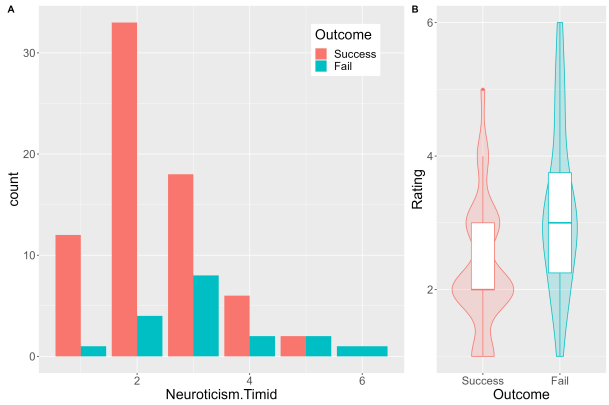



Figure B.24: Bar chart and box plot of the distribution of the MCPQ-R item "Neuroticism Timid" in relation to training outcomes.

# Appendix C

# Behavioural Assessment - Test Battery

## C.1 Ethogram

Table C.1: Detailed description of the ethogram for each of the sub-tests in the novel Assistance Dog Test Battery (ADTB), including the behavioural items analysed and the marking system.

| Items | Marking |
|---|---|
| **1. Familiarisation** | |
| Oriented to Handler;<br>Exploration;<br>Waiting<br>(Harvey et al. 2016b; Palestrini et al. 2005) | 1 - Never;<br>2 - Rarely;<br>3 - Occasionally;<br>4 - Regularly;<br>5 - Constantly |
| Call Back-Response | 1 - On the first command;<br>2 - On the second or more;<br>3 - Did not respond accordingly |
| **2. Walking** | |
| Distractibility | 1 - Hardly distracted (1-2 times);<br>2 - Somewhat distracted (3-4 times);<br>3 - Mostly distracted (5+ times) |
| Pull on leash<br>(Harvey et al. 2016b) | 1 - Not observed;<br>2 - Rarely;<br>3 - Occasionally;<br>4 - Regularly;<br>5 - Constantly |

Continued on next page.

Table C.1 - continued from previous page.

| Items | Marking |
|---|---|
| Pull strength | 1 - No tension, lead relaxed;<br>2 - Mild tension;<br>3 - Moderate tension;<br>4 - Strong tension;<br>5 - Extreme tension, back legs stretched and on or more front paws raised off the floor |
| Initiative<br>(Bergamasco et al. 2010) | 1 - Dog walks out ahead the handler;<br>2 - Dog walks beside/at heel with the handlers;<br>3 - Dog walks occasionally behind the handler, needs encouragements;<br>4 - Dog walks consistently behind the handler;<br>5 - Dog refuses to walk at times |
| **3. Standing** | |
| Response<br>(Harvey et al. 2016b) | 1 - Poor (5+ attempts);<br>2 - Not great (3-4 attempts);<br>3 - Good (2 attempts);<br>4 - Excellent (1 attempt) |
| **4. Siting** | |
| Response<br>(Harvey et al. 2016b) | 1 - Poor (5+ attempts);<br>2 - Not great (3-4 attempts);<br>3 - Good (2 attempts);<br>4 - Excellent (1 attempt) |
| **5. Lying** | |
| Response | 1 - Poor (5+ attempts);<br>2 - Not great (3-4 attempts);<br>3 - Good (2 attempts);<br>4 - Excellent (1 attempt) |
| Settled | 1 - Very unsettled (high level of distraction, high level of handler support);<br>2 - Slightly unsettled (medium level of distraction, moderate handler support);<br>3 - Reasonably settled (low level of distraction, minimum handler support);<br>4 - Very settled (no handler support) |
| **6. Body check** | |
| Table | 1 - Refuse;<br>2 - Hesitant;<br>3 - Cooperative;<br>4 - Willing |
| Response | 1 - Escaping/Not willing;<br>2 - Reluctant;<br>3 - Cooperative;<br>4 - Willing |

Continued on next page.

Table C.1 - continued from previous page.

| Items | Marking |
|---|---|
| Behaviours<br>Mouths;<br>Licks<br>(Harvey et al. 2016b) | Yes;<br>No |
| **7. Tea towel** | |
| [First Response]<br>Indifferent;<br>Changes from Neutral;<br>Turn head;<br>Attempts to/Removes towel by moving;<br>Attempts to/Removes towel with mouth;<br>Plays<br>(Harvey et al. 2016b; Palestrini et al. 2005) | Yes;<br>No |
| [Second Response]<br>Indifferent;<br>Changes from Neutral;<br>Turn head;<br>Attempts to/Removes towel by moving;<br>Attempts to/Removes towel with mouth;<br>Plays<br>(Harvey et al. 2016b; Palestrini et al. 2005) | Yes;<br>No |
| **8. Distractions** | |
| [First Response]<br>Teddy;<br>Human;<br>Car;<br>Food<br>(Harvey et al. 2016b; Foyer et al. 2016) | 1 - Does not notice;<br>2 - Notices, does not interact;<br>3 - Notices, interacts, moves on quickly;<br>4 - Notices, interacts, moves on after a pause;<br>5 - Notices, interacts, moves on after some time;<br>6 - Notices, interacts, does not move on without serious intervention |
| [Second Response]<br>Teddy;<br>Human;<br>Car;<br>Food<br>(Harvey et al. 2016b; Foyer et al. 2016) | 1 - Does not notice;<br>2 - Notices, does not interact;<br>3 - Notices, interacts, moves on quickly;<br>4 - Notices, interacts, moves on after a pause;<br>5 - Notices, interacts, moves on after some time;<br>6 - Notices, interacts, does not move on without serious intervention |

Continued on next page.

Table C.1 - continued from previous page.

| Items | Marking |
|---|---|
| Pull on leash (Harvey et al. 2016b) | 1 - Not observed (0) ;<br>2 - Rare (1-2);<br>3 - Sometimes (3-5);<br>4 - Frequent (6-8) |
| Pull strength (Harvey et al. 2016b) | 1 - No tension, lead relaxed;<br>2 - Mild tension;<br>3 - Moderate tension;<br>4 - Strong tension;<br>5 - Extreme tension, back legs stretched and on or more front paws raised off the floor |
| Sociability (Bergamasco et al. 2010; Foyer et al. 2016) | 1 - Runs away or is afraid;<br>2 - Looks for the handler or is hesitant toward the person;<br>3 - Is indifferent;<br>4 - Positively looks at person, does not engage with them;<br>5 - Approaches the person;<br>6 - Is very friendly, licks the person's hands and wags its tail;<br>7 - Gets overly excited, jumps, mouths person |
| | **9. Kong** |
| Presentation Response | 1 - Not interested, looks or walks away;<br>2 - Somewhat interested;<br>3 - Mostly interested;<br>4 - Very interested;<br>5 - Overly interested, dog jumps at handler |
| Interaction Lateralisation | None;<br>Left;<br>Right;<br>Both |
| Interaction Response to Stimulus (Asher et al. 2013) | 1 - Does not interact with Kong;<br>2 - Interacts with Kong, but loses interest;<br>3 - Interacts with Kong continuously, but calmly;<br>4 - Interacts with Kong continuously and intensely |
| Interaction Response to Handler (Asher et al. 2013) | 1 - Stays close to handler;<br>2 - Moves slightly away from handler;<br>3 - Takes Kong and move away from handler |

Continued on next page.

Table C.1 - continued from previous page.

| Items | Marking |
|---|---|
| Return Response to Stimulus | 1 - Not bothered to let it go;<br>2 - Mildly reluctant, but lets it go;<br>3 - Moderately reluctant, but lets it go;<br>4 - Very reluctant, does not let it go;<br>5 - Extremely reluctant, does not let it go and tries to get it back |
| Return Response to Handler | 1 - Does not follow the handler;<br>2 - Calmly follows the handler;<br>3 - Actively follows the handler;<br>4 - Intensely follows the handler |
| **10. Dog distraction** | |
| Response | 1 - Cautious, seeks handler support;<br>2 - Not interest in other dog, does not move;<br>3 - Not interested in other dog, explores environment;<br>4 - Mildly interested in other dog, does not move;<br>5 - Moderately interested in other dog, moves towards dog after sometime;<br>6 - Very interested, moves towards dog immediately;<br>7 - Very excited, runs after dog instantly;<br>8 - Extremely excited, runs after dog instantly and plays around a lot |
| Call Back Response | 1 - Does not leave trainer's side/comes back without being called;<br>2 - On the first command;<br>3 - On the second command or more;<br>4 - Does not respond accordingly |
| **11. Crate** | |
| Entering | 1 - Without any encouragement;<br>2 - Needs some encouragement;<br>3 - Needs a lot of encouragement;<br>4 - Won't go in, almost forced/did not go |
| Settled;<br>Sniffing/Exploring;<br>Actively Seeking Attention;<br>Digging;<br>Nudging;<br>Pawing;<br>Whining;<br>Barking | 1 - Never;<br>2 - Rarely;<br>3 - Occasionally;<br>4 - Regularly;<br>5 - Constantly |
| Self-Modulation | 1 - Good;<br>2 - Slow;<br>3 - Poor |

Continued on next page.

Table C.1 - continued from previous page.

| Items | Marking |
|---|---|
| Stress | 1 - Settled;<br>2 - Mildly stressed;<br>3 - Moderately stressed;<br>4 - Extremely stressed |
| Urinating | Yes;<br>No |
| **12. Petting** | |
| Handler-Stimulus | 1 - Slightly playful;<br>2 - Moderately playful;<br>3 - Extremely playful |
| Handler-Holding dog | 1 - Never;<br>2 - Sometimes;<br>3 - Always |
| Confidence-During<br>(Asher et al. 2013) | 1 - Avoids, moves away;<br>2 - Indifferent;<br>3 - Slightly excited, tail wagging;<br>4 - Very excited, moving around;<br>5 - Extremely excited, jumping up, licking, mouthing |
| Responsiveness-After<br>(Asher et al. 2013) | 1 - Anxious;<br>2 - Indifferent;<br>3 - Explorative;<br>4 - Calmly interacts with the handler;<br>5 - Mild attention seeking;<br>6 - Moderate attention seeking |
| **13. Isolation** | |
| Time oriented;<br>Exploration;<br>Unsettled/Pacing;<br>Whining;<br>Barking<br>(Harvey et al. 2016b; Foyer et al. 2016; Katayama et al. 2016; Palestrini et al. 2005) | 1 - Never;<br>2 - Rarely;<br>3 - Occasionally;<br>4 - Regularly;<br>5 - Constantly |
| Urinating | Yes;<br>No |
| Reunion<br>(Harvey et al. 2016b) | 1 - Runs away;<br>2 - Indifferent;<br>3 - Explorative;<br>4 - Calm approach;<br>5 - Activated Approach;<br>6 - Overly Activated Approach |
| **14. Noise** | |

Continued on next page.

Table C.1 - continued from previous page.

| Items | Marking |
|---|---|
| Confidence (Asher et al. 2013) | 1 - Runs away, shakes;<br>2 - Moves away, avoids;<br>3 - Indifferent;<br>4 - Minor reaction;<br>5 - Moderate reaction;<br>6 - Strong reaction (jumps, runs to/paws at the source of the sound, vocalises) |

Continued on next page.

All variables had an NA option
The variables recorded were classified into ordinal variables and categorical variables depending on whether numbers were used or not: Ordinal variables - all markings with numbers; Categorical variables - all markings without number

## C.2 Composite scores

In order to combine information from multiple ethogram original items into one variable, composite scores were derived from a number of original items. Table C.2 describes the calculation of the 41 composite scores in more detail. Ordinal variables were combined by calculating both their mean and product. Categorical variables were combined by assigning a numerical order to them.

Table C.2: Composite scores originated from Assistance Dog Test Battery (ADTB) ethogram items presented in Table C.1. For numerical ethogram items, composite scores were calculated by taking the mean and product between them. For categorical ethogram items, composite scores were derived by assigning numbers to categories in a meaningful way.

| Composite scores (calculated) | Ethogram items (original) |
|---|---|
| **1. Familiarisation** | |
| S Familiarisation Handler_mean<br>S Familiarisation Handler_prod | Oriented to Handler<br>Waiting |
| **2. Walking** | |
| S Walking Pull_mean<br>S Walking Pull_prod | Walking Pull on leash<br>Walking Pull strength |
| **3-5. Standing, Sitting, Lying** | |
| S Obedience_mean<br>S Obedience_prod | Standing Response<br>Sitting Response<br>Lying Response |
| **6. Body check** | |
| S Sensitivity_mean | Table<br>Response |

Continued on next page

Table C.2 - continued from previous page

| Composite scores (calculated) | Ethogram items (original) |
|---|---|
| **7. Tea towel** | |
| S Tea Towel First Response | (Mean for [First Response] of the following)<br>Indifferent (replace: Yes = 1, No = 0)<br>Change from Neutral (replace: Yes = 2, No = 0)<br>Turns head (replace: Yes = 3, No = 0)<br>Attempts to/Removes towel by moving (replace: Yes = 4, No = 0)<br>Attempts to/Removes towel with mouth (replace Yes = 5, No = 0)<br>Plays (replace: Yes = 6, No = 0) |
| S Tea Towel Second Response | (Mean for [Second Response] of the following)<br>Indifferent (replace: Yes = 1, No = 0)<br>Change from Neutral (replace: Yes = 2, No = 0)<br>Turns head (replace: Yes = 3, No = 0)<br>Attempts to/Removes towel by moving (replace: Yes = 4, No = 0)<br>Attempts to/Removes towel with mouth (replace Yes = 5, No = 0)<br>Plays (replace: Yes = 6, No = 0) |
| **8. Distractions** | |
| S Distractions First Response_mean<br>S Distractions First Response_prod | [First Response]<br>Teddy<br>Human<br>Car<br>Food |
| S Distractions Second Response_mean<br>S Distractions Second Response_prod | [Second Response]<br>Teddy<br>Human<br>Car<br>Food |
| S Distractions Difference_mean | (Mean of the differences between individual responses, i.e.: [Second Response] [First Response])<br>Teddy<br>Human<br>Car<br>Food |
| S Distractions Pull_mean<br>S Distractions Pull_prod | Distractions Pull on leash<br>Distractions Pull strength |
| S Walking Distractions Pull_mean | S Distractions Pull_mean<br>S Walking Pull_mean |
| S Walking Distractions Pull_prod | S Distractions Pull_prod<br>S Walking Pull_prod |

Continued on next page

Table C.2 - continued from previous page

| Composite scores (calculated) | Ethogram items (original) |
|---|---|
| | **9. Kong** |
| S Kong Response_mean<br>S Kong Response_prod | Presentation Response<br>Interaction Response to stimulus<br>Interaction Back<br>Return Handler |
| | **10. Dog distraction** |
| S Dog Distraction_mean<br>S Dog Distraction_prod | Response<br>Call Back Response |
| S Call Response_mean<br>S Call Response_prod | Familiarisation Call Back Response<br>Dog Call Back Response |
| S Familiarisation Dog Call | (Difference between the following)<br>Familiarisation Call Back Response<br>Dog Call Back Response |
| | **11. Crate** |
| S Crate Response_mean<br>S Crate Response_prod | Entering<br>Self-Modulation<br>Stress<br>Settled (in reverse order) |
| S Crate Stimulus_mean<br>S Crate Stimulus_prod | Sniffing/Exploring<br>Digging<br>Nudging Crate<br>Pawing |
| S Crate Handler_mean<br>S Crate Handler_prod | Actively Seeking Attention<br>Whining<br>Barking |
| | **12. Petting** |
| S Petting Stimulus During_mean<br>S Petting Stimulus During_prod | Handler Stimulus<br>Confidence During |
| S Petting Engagement During_mean<br>S Petting Engagement During_prod | Handler Holding dog<br>Confidence During |
| S Petting After_mean<br>S Petting After_prod | Handler Holding dog<br>Responsiveness After |
| S Sociability_mean<br>S Sociability_prod | Confidence During<br>Responsiveness After<br>Reunion Response |
| | **13. Isolation** |
| S Isolation Handler_mean | Time oriented<br>Whining<br>Barking |
| S Isolation Stimulus | (Difference between the following)<br>Exploration<br>Unsettled/Pacing |

Continued on next page

# C.3 Data Collection 1 (DC1)

## C.3.1 Descriptive Statistics

Table C.1 shows the ethogram used to mark the behaviours shown in each subtest. Table C.3 shows descriptive statistics of the ordinal ethogram items and composite scores of the data from DC1, including the minimum, 1st quartile, median, mean, 3rd quartile, maximum and number of missing values for all ethogram items and scores. Table C.4 shows descriptive statistics of the categorical variables including the number of dogs who scored “Yes”, or “No”, or are missing the score. The only categorical variable which has different categories is “Kong - Interaction Lateralisation”, therefore the descriptive statistics are shown in Table C.5.

Table C.3: Descriptive statistics for DC2 numerical variables of the ADTB ethogram including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing values for all items and scores.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs |
|---|---|---|---|---|---|---|---|
| Familiarisation Response Oriented to Handler | 1.00 | 1.50 | 2.00 | 2.24 | 3.00 | 5.00 | 3 (0.03) |
| Familiarisation Response Exploration | 1.00 | 4.00 | 5.00 | 4.29 | 5.00 | 5.00 | 0 (0.00) |
| Familiarisation Response Waiting | 1.00 | 1.00 | 1.00 | 1.52 | 2.00 | 5.00 | 1 (0.01) |
| Call Back Response | 1.00 | 1.00 | 1.00 | 1.33 | 2.00 | 3.00 | 15 (0.15) |
| Walking Distractibility | 1.00 | 1.00 | 2.00 | 1.92 | 3.00 | 3.00 | 0 (0.00) |
| Walking Pull on leash | 1.00 | 1.00 | 1.50 | 1.89 | 3.00 | 4.00 | 0 (0.00) |
| Walking Pull strength | 1.00 | 1.00 | 1.00 | 1.62 | 2.00 | 4.00 | 0 (0.00) |
| Walking Initiative | 1.00 | 2.00 | 2.00 | 2.42 | 3.00 | 5.00 | 0 (0.00) |
| Standing Response | 1.00 | 2.00 | 2.00 | 2.48 | 3.00 | 4.00 | 0 (0.00) |
| Sitting Response | 1.00 | 2.00 | 3.00 | 2.56 | 3.25 | 4.00 | 2 (0.02) |
| Lying Response | 1.00 | 2.00 | 3.00 | 2.66 | 4.00 | 4.00 | 0 (0.00) |
| Lying Settled | 1.00 | 1.00 | 2.00 | 2.29 | 3.00 | 4.00 | 2 (0.02) |
| Body check Table | 1.00 | 3.00 | 4.00 | 3.30 | 4.00 | 4.00 | 5 (0.05) |
| Body check Response | 1.00 | 3.00 | 3.00 | 2.93 | 3.00 | 4.00 | 0 (0.00) |
| Distractions First Response Teddy | 1.00 | 3.00 | 4.00 | 3.67 | 4.75 | 6.00 | 0 (0.00) |
| Distractions First Response Human | 2.00 | 3.00 | 3.00 | 3.36 | 4.00 | 5.00 | 0 (0.00) |
| Distractions First Response Car | 1.00 | 2.25 | 3.00 | 3.12 | 4.00 | 6.00 | 0 (0.00) |
| Distractions First Response Food | 1.00 | 4.00 | 4.00 | 4.22 | 5.00 | 6.00 | 0 (0.00) |
| Distractions Second Response Teddy | 1.00 | 2.00 | 3.00 | 2.69 | 3.00 | 6.00 | 0 (0.00) |
| Distractions Second Response Human | 1.00 | 2.00 | 3.00 | 2.95 | 4.00 | 5.00 | 1 (0.01) |
| Distractions Second Response Car | 1.00 | 2.00 | 3.00 | 2.61 | 3.00 | 5.00 | 1 (0.01) |
| Distractions Second Response Food | 1.00 | 3.00 | 4.00 | 3.79 | 5.00 | 6.00 | 2 (0.02) |
| Distractions Pull on leash | 1.00 | 3.00 | 3.00 | 3.00 | 3.75 | 4.00 | 0 (0.00) |

Continued on next page.

Table C.3 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| Distractions Pull strength | 1.00 | 3.00 | 3.00 | 3.52 | 5.00 | 5.00 | 0 (0.00) |
| Distractions Human | 3.00 | 4.00 | 5.00 | 5.24 | 6.00 | 7.00 | 0 (0.00) |
| Kong Presentation Response | 1.00 | 2.00 | 4.00 | 3.52 | 5.00 | 5.00 | 0 (0.00) |
| Kong Interaction Response to stimulus | 1.00 | 3.00 | 3.00 | 2.95 | 3.00 | 4.00 | 1 (0.01) |
| Kong Interaction Response to handler | 1.00 | 1.00 | 1.00 | 1.55 | 2.00 | 3.00 | 2 (0.02) |
| Kong Interaction Back | 1.00 | 1.00 | 1.00 | 1.56 | 2.00 | 5.00 | 4 (0.04) |
| Kong Return Handler | 1.00 | 2.00 | 2.00 | 2.14 | 2.00 | 4.00 | 4 (0.04) |
| Dog Response | 1.00 | 4.00 | 5.00 | 4.91 | 6.00 | 8.00 | 0 (0.00) |
| Dog Call Back Response | 1.00 | 1.00 | 2.00 | 1.83 | 2.00 | 4.00 | 5 (0.05) |
| Crate Entering | 1.00 | 1.00 | 1.00 | 1.44 | 2.00 | 4.00 | 1 (0.01) |
| Crate Settled | 1.00 | 1.00 | 2.00 | 2.23 | 3.00 | 5.00 | 2 (0.02) |
| Crate Sniffing Exploring | 1.00 | 2.00 | 3.00 | 2.72 | 3.00 | 5.00 | 2 (0.02) |
| Crate Actively Seeking Attention | 1.00 | 1.00 | 1.00 | 2.06 | 3.00 | 5.00 | 3 (0.03) |
| Crate Digging | 1.00 | 1.00 | 1.00 | 1.14 | 1.00 | 4.00 | 2 (0.02) |
| Crate Nudging Crate | 1.00 | 1.00 | 1.00 | 1.18 | 1.00 | 3.00 | 2 (0.02) |
| Crate Pawing | 1.00 | 1.00 | 1.00 | 1.26 | 1.00 | 4.00 | 2 (0.02) |
| Crate Whining | 1.00 | 1.00 | 1.00 | 1.89 | 3.00 | 5.00 | 8 (0.08) |
| Crate Barking | 1.00 | 1.00 | 1.00 | 1.48 | 1.00 | 5.00 | 3 (0.03) |
| Crate Self Modulation | 1.00 | 1.00 | 2.00 | 2.06 | 3.00 | 3.00 | 2 (0.02) |
| Crate Stress | 1.00 | 2.00 | 2.00 | 2.30 | 3.00 | 4.00 | 2 (0.02) |
| Petting Handler Stimulus | 1.00 | 1.00 | 1.00 | 1.18 | 1.00 | 3.00 | 1 (0.01) |
| Petting Handler Holding dog | 1.00 | 1.00 | 1.00 | 1.16 | 1.00 | 3.00 | 1 (0.01) |
| Petting Confidence During | 1.00 | 3.00 | 3.00 | 3.04 | 3.00 | 5.00 | 1 (0.01) |
| Petting Responsiveness After | 1.00 | 3.00 | 4.00 | 3.76 | 4.00 | 6.00 | 2 (0.02) |
| Isolation Response Time oriented | 1.00 | 3.00 | 4.00 | 3.94 | 5.00 | 5.00 | 2 (0.02) |

Continued on next page.

Table C.3 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| Isolation Response Exploration | 1.00 | 1.00 | 2.00 | 2.09 | 3.00 | 5.00 | 1 (0.01) |
| Isolation Response Unsettled Pacing | 1.00 | 1.00 | 2.00 | 1.96 | 2.00 | 5.00 | 1 (0.01) |
| Isolation Response Whining | 1.00 | 1.00 | 2.00 | 2.02 | 3.00 | 5.00 | 8 (0.08) |
| Isolation Response Barking | 1.00 | 1.00 | 1.00 | 1.46 | 1.00 | 5.00 | 1 (0.01) |
| Isolation Reunion | 1.00 | 4.00 | 4.00 | 4.20 | 5.00 | 6.00 | 2 (0.02) |
| Noise Confidence | 3.00 | 4.00 | 4.00 | 3.88 | 4.00 | 6.00 | 1 (0.01) |
| S Familiarisation Handler_mean | 1.00 | 1.00 | 1.50 | 1.87 | 2.50 | 5.00 | 1 (0.01) |
| S Familiarisation Handler_prod | 1.00 | 1.00 | 2.00 | 3.83 | 4.00 | 25.00 | 0 (0.00) |
| S Walking Pull_mean | 1.00 | 1.00 | 1.50 | 1.76 | 2.50 | 4.00 | 0 (0.00) |
| S Walking Pull_prod | 1.00 | 1.00 | 2.00 | 3.72 | 6.00 | 16.00 | 0 (0.00) |
| S Distractions Pull_mean | 1.00 | 3.00 | 3.00 | 3.26 | 4.00 | 4.50 | 0 (0.00) |
| S Distractions Pull_prod | 1.00 | 9.00 | 9.00 | 11.15 | 15.00 | 20.00 | 0 (0.00) |
| S Walking Distractions Pull_mean | -2.00 | 1.00 | 1.50 | 1.51 | 2.00 | 3.50 | 0 (0.00) |
| S Walking Distractions Pull_prod | -12.00 | 3.00 | 8.00 | 7.43 | 11.00 | 19.00 | 0 (0.00) |
| S Obedience_mean | 1.00 | 2.00 | 2.67 | 2.56 | 3.00 | 4.00 | 0 (0.00) |
| S Obedience_prod | 1.00 | 6.50 | 12.00 | 17.58 | 24.00 | 64.00 | 0 (0.00) |
| S Sensitivity_mean | 1.50 | 3.00 | 3.00 | 3.11 | 3.50 | 4.00 | 0 (0.00) |
| S Tea Towel First Response | 0.00 | 1.00 | 1.67 | 1.68 | 2.50 | 3.50 | 1 (0.01) |
| S Tea Towel Second Response | 0.00 | 0.50 | 1.67 | 1.48 | 2.50 | 3.50 | 3 (0.03) |
| S Distractions First Response_mean | 2.00 | 3.00 | 3.50 | 3.59 | 4.25 | 5.50 | 0 (0.00) |
| S Distractions First Response_prod | 8.00 | 61.00 | 144.00 | 190.66 | 256.00 | 900.00 | 0 (0.00) |
| S Distractions Second Response_mean | 1.50 | 2.50 | 3.00 | 3.02 | 3.50 | 5.50 | 0 (0.00) |
| S Distractions Second Response_prod | 3.00 | 24.00 | 54.00 | 98.14 | 108.00 | 600.00 | 0 (0.00) |
| S Distractions Difference_mean | -2.50 | -1.00 | -0.58 | -0.58 | -0.25 | 1.75 | 0 (0.00) |
| S Kong Response_mean | 1.33 | 2.00 | 2.50 | 2.55 | 2.75 | 4.25 | 0 (0.00) |

Continued on next page.

Table C.3 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| S Kong Response_prod | 2.00 | 12.00 | 24.00 | 39.59 | 48.00 | 300.00 | 0 (0.00) |
| S Dog Distraction_mean | 1.00 | 2.50 | 3.50 | 3.45 | 4.00 | 6.00 | 0 (0.00) |
| S Dog Distraction_prod | 1.00 | 4.00 | 6.00 | 9.27 | 12.00 | 32.00 | 0 (0.00) |
| S Call Response_mean | 1.00 | 1.00 | 1.50 | 1.59 | 2.00 | 3.00 | 2 (0.02) |
| S Call Response_prod | 1.00 | 1.00 | 2.00 | 2.35 | 3.00 | 8.00 | 0 (0.00) |
| S Familiarisation Dog Call | -3.00 | -1.00 | 0.00 | -0.56 | 0.00 | 1.00 | 18 (0.18) |
| S Crate Response_mean | 0.75 | 1.75 | 2.25 | 2.16 | 2.75 | 4.00 | 1 (0.01) |
| S Crate Response_prod | 0.00 | 4.00 | 16.00 | 25.19 | 27.00 | 144.00 | 0 (0.00) |
| S Crate Stimulus_mean | 1.25 | 1.25 | 1.50 | 1.57 | 1.75 | 3.25 | 2 (0.02) |
| S Crate Stimulus_prod | 1.00 | 2.00 | 3.00 | 6.24 | 4.00 | 108.00 | 0 (0.00) |
| S Crate Handler_mean | 1.00 | 1.00 | 1.33 | 1.80 | 2.42 | 4.00 | 2 (0.02) |
| S Crate Handler_prod | 1.00 | 1.00 | 2.00 | 9.27 | 9.00 | 60.00 | 0 (0.00) |
| S Petting Stimulus During_mean | 1.00 | 2.00 | 2.00 | 2.11 | 2.50 | 4.00 | 1 (0.01) |
| S Petting Stimulus During_prod | 1.00 | 3.00 | 3.00 | 3.67 | 4.00 | 15.00 | 0 (0.00) |
| S Petting Engagement During_mean | 1.00 | 2.00 | 2.00 | 2.10 | 2.50 | 3.50 | 1 (0.01) |
| S Petting Engagement During_prod | 1.00 | 3.00 | 3.00 | 3.49 | 4.00 | 10.00 | 0 (0.00) |
| S Petting After_mean | 1.00 | 2.00 | 2.50 | 2.45 | 2.50 | 3.50 | 1 (0.01) |
| S Petting After_prod | 1.00 | 3.00 | 4.00 | 4.26 | 4.00 | 12.00 | 0 (0.00) |
| S Isolation Handler_mean | 1.00 | 2.00 | 2.33 | 2.47 | 3.00 | 4.33 | 1 (0.01) |
| S Isolation Handler_prod | 1.00 | 4.00 | 8.00 | 12.24 | 15.75 | 75.00 | 0 (0.00) |
| S Isolation Stimulus | -4.00 | -1.00 | 0.00 | 0.13 | 1.00 | 3.00 | 1 (0.01) |
| S Sociability_mean | 2.00 | 3.33 | 3.67 | 3.67 | 4.00 | 5.33 | 0 (0.00) |
| S Sociability_prod | 4.00 | 36.00 | 48.00 | 49.07 | 60.00 | 150.00 | 0 (0.00) |

Table C.4: Descriptive statistics of DC1 for categorical variables of the ADTB ethogram including the number of No, Yes and missing data.

| Variable | No | Yes | NAs (%) |
|---|---|---|---|
| Body check General Mouths | 94 | 3 | 1 (0.01) |
| Body check General Licks | 84 | 13 | 1 (0.01) |
| Tea Towel First Response Indifferent | 87 | 10 | 1 (0.01) |
| Tea Towel First Response Change from Neutral | 13 | 84 | 1 (0.01) |
| Tea Towel First Response Turns head | 17 | 80 | 1 (0.01) |
| Tea Towel First Response Attempts to Removes towel by moving | 48 | 49 | 1 (0.01) |
| Tea Towel First Response Attempts to Removes towel with mouth | 52 | 45 | 1 (0.01) |
| Tea Towel First Response Plays | 87 | 10 | 1 (0.01) |
| Tea Towel Second Response Indifferent | 71 | 24 | 3 (0.03) |
| Tea Towel Second Response Turns head | 31 | 64 | 3 (0.03) |
| Tea Towel Second Response Change from Neutral | 26 | 69 | 3 (0.03) |
| Tea Towel Second Response Attempts to Removes towel by moving | 50 | 45 | 3 (0.03) |
| Tea Towel Second Response Attempts to Removes towel with mouth | 58 | 37 | 3 (0.03) |
| Tea Towel Second Response Plays | 82 | 13 | 3 (0.03) |
| Crate Urinating | 95 | 1 | 2 (0.02) |
| Isolation Urinating | 97 | 1 | 1 (0.01) |

Table C.5: ADTB ethogram descriptive statistics for the variable Kong Interaction Lateralisation, including the number of dogs in the sample per category.

| Variable | None | Both | Right | Left |
|---|---|---|---|---|
| Kong - Interaction Lateralisation | 43 | 10 | 22 | 17 |

Table C.6: Descriptive statistics of the composite scores used in the optimal multivariate logistic regression model and the items from which they derive.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs |
|---|---|---|---|---|---|---|---|
| **S Sensitivity_mean** | 1.50 | 3.00 | 3.00 | 3.11 | 3.50 | 4.00 | 0 |
| Body check Table | 1.00 | 3.00 | 4.00 | 3.30 | 4.00 | 4.00 | 5 |
| Body check Response | 1.00 | 3.00 | 3.00 | 2.93 | 3.00 | 4.00 | 0 |
| **S Distractions Pull_mean** | 1.00 | 3.00 | 3.00 | 3.26 | 4.00 | 4.50 | 0 |
| Distractions Pull on leash | 1.00 | 3.00 | 3.00 | 3.00 | 3.75 | 4.00 | 0 |
| Distractions Pull strength | 1.00 | 3.00 | 3.00 | 3.52 | 5.00 | 5.00 | 0 |
| **S Crate Stimulus_mean** | 1.25 | 1.25 | 1.50 | 1.57 | 1.75 | 3.25 | 2 |
| Crate Sniffing Exploring | 1.00 | 2.00 | 3.00 | 2.72 | 3.00 | 5.00 | 2 |
| Crate Digging | 1.00 | 1.00 | 1.00 | 1.14 | 1.00 | 4.00 | 2 |
| Crate Nudging Crate | 1.00 | 1.00 | 1.00 | 1.18 | 1.00 | 3.00 | 2 |
| Crate Pawing | 1.00 | 1.00 | 1.00 | 1.26 | 1.00 | 4.00 | 2 |
| **S Petting Stimulus During_mean** | 1.00 | 2.00 | 2.00 | 2.11 | 2.50 | 4.00 | 1 |
| Petting Handler Stimulus | 1.00 | 1.00 | 1.00 | 1.18 | 1.00 | 3.00 | 1 |
| Petting Confidence During | 1.00 | 3.00 | 3.00 | 3.04 | 3.00 | 5.00 | 1 |

### C.3.2 Univariate Analysis

Multiple models were built using each ethogram variable individually to predict the probability of withdrawal from training. The results from the univariate logistic regression models are shown in C.7. It includes the model estimate, p-value, Odds Ratio, and the lower 2.5% and upper 97.5% bounds of the Confidence Interval (CI). Positive estimates indicate that the variable was positively associated with a higher probability of withdrawal. Negative estimates suggest that the variable was negatively

associated with a higher probability of withdrawal. The p-value shows the level of significance of this relationship.

Table C.7: ADTB univariate logistic regression model estimates, p-value, Odds Ratio (OR), and Confidence Interval (CI).

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| Items - Ordinal | | | | | |
| Familiarisation Response Oriented to Handler | 0.24 | 0.27 | 1.28 | 0.82 | 1.96 |
| Familiarisation Response Exploration | -0.32 | 0.15 | 0.73 | 0.47 | 1.14 |
| Familiarisation Response Waiting | 0.20 | 0.50 | 1.22 | 0.66 | 2.13 |
| Call Back Response | 0.20 | 0.68 | 1.22 | 0.43 | 3.11 |
| Walking Distractibility | 0.07 | 0.82 | 1.07 | 0.58 | 1.98 |
| Walking Pull on leash | -0.10 | 0.69 | 0.91 | 0.54 | 1.45 |
| Walking Pull strength | -0.12 | 0.72 | 0.88 | 0.43 | 1.69 |
| Walking Initiative | 0.45 | 0.17 | 1.56 | 0.83 | 3.00 |
| Standing Response | 0.11 | 0.65 | 1.12 | 0.70 | 1.80 |
| Sitting Response | 0.01 | 0.97 | 1.01 | 0.64 | 1.59 |
| Lying Response | 0.05 | 0.82 | 1.05 | 0.69 | 1.62 |
| Lying Settled | 0.20 | 0.43 | 1.22 | 0.75 | 2.00 |
| Body check Table ** | -0.48 | 0.07 | 0.62 | 0.37 | 1.05 |
| Body check Response | -0.19 | 0.59 | 0.83 | 0.42 | 1.69 |
| Distractions First Response Teddy | -0.09 | 0.59 | 0.91 | 0.65 | 1.28 |
| Distractions First Response Human * | -0.40 | 0.15 | 0.67 | 0.38 | 1.14 |
| Distractions First Response Car | -0.30 | 0.25 | 0.74 | 0.44 | 1.22 |
| Distractions First Response Food | 0.12 | 0.54 | 1.13 | 0.78 | 1.71 |
| Distractions Second Response Teddy | -0.02 | 0.92 | 0.98 | 0.68 | 1.39 |
| Distractions Second Response Human | -0.38 | 0.19 | 0.69 | 0.38 | 1.18 |
| Distractions Second Response Car | 0.02 | 0.95 | 1.02 | 0.60 | 1.69 |
| Distractions Second Response Food | 0.20 | 0.31 | 1.22 | 0.84 | 1.81 |
| Distractions Pull on leash | -0.44 | 0.19 | 0.65 | 0.33 | 1.24 |

Continued on next page.

Table C.7 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| Distractions Human | -0.34 | 0.19 | 0.71 | 0.42 | 1.18 |
| Kong Presentation Response | 0.11 | 0.57 | 1.12 | 0.77 | 1.64 |
| Kong Interaction Response to stimulus | 0.01 | 0.99 | 1.01 | 0.40 | 2.71 |
| Kong Interaction Response to handler | 0.03 | 0.91 | 1.03 | 0.57 | 1.79 |
| Kong Interaction Back | 0.22 | 0.34 | 1.25 | 0.76 | 1.97 |
| Kong Return Handler | 0.40 | 0.47 | 1.49 | 0.47 | 4.35 |
| Dog Response | -0.11 | 0.56 | 0.90 | 0.61 | 1.30 |
| Dog Call Back Response | 0.02 | 0.94 | 1.02 | 0.58 | 1.74 |
| Crate Entering | 0.16 | 0.67 | 1.17 | 0.54 | 2.36 |
| Crate Behaviours Settled | 0.03 | 0.89 | 1.03 | 0.66 | 1.57 |
| Crate Behaviours Sniffing Exploring ** | 0.59 | 0.06 | 1.80 | 0.99 | 3.37 |
| Crate Behaviours Actively Seeking Attention * | -0.37 | 0.12 | 0.69 | 0.41 | 1.06 |
| Crate Behaviours Digging | 0.48 | 0.25 | 1.62 | 0.66 | 3.71 |
| Crate Behaviours Whining | -0.19 | 0.47 | 0.83 | 0.47 | 1.34 |
| Crate Behaviours Barking | -0.31 | 0.32 | 0.73 | 0.34 | 1.24 |
| Crate Self Modulation | -0.22 | 0.49 | 0.81 | 0.43 | 1.50 |
| Crate Behaviours Nudging Crate | 0.57 | 0.19 | 1.77 | 0.71 | 4.15 |
| Crate Behaviours Pawing | 0.01 | 0.98 | 1.01 | 0.40 | 2.08 |
| Crate Stress | -0.21 | 0.55 | 0.81 | 0.40 | 1.59 |
| Petting Handler Stimulus * | -1.56 | 0.14 | 0.21 | 0.01 | 1.08 |
| Petting Handler Holding dog | 0.73 | 0.15 | 2.07 | 0.73 | 5.73 |
| Petting Confidence During *** | -0.66 | 0.03 | 0.51 | 0.28 | 0.91 |
| Petting Responsiveness After | -0.41 | 0.17 | 0.67 | 0.37 | 1.17 |
| Isolation Response Time oriented | -0.05 | 0.85 | 0.95 | 0.56 | 1.65 |
| Isolation Response Exploration | 0.06 | 0.80 | 1.06 | 0.66 | 1.68 |

Continued on next page.

Table C.7 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| Isolation Response Unsettled Pacing | 0.12 | 0.63 | 1.13 | 0.67 | 1.82 |
| Isolation Response Whining | -0.17 | 0.53 | 0.85 | 0.49 | 1.40 |
| Isolation Response Barking | 0.07 | 0.76 | 1.08 | 0.64 | 1.68 |
| Isolation Reunion | 0.00 | 0.99 | 1.00 | 0.57 | 1.86 |
| Noise Confidence | -0.20 | 0.60 | 0.82 | 0.36 | 1.69 |
| Scores | | | | | |
| S Familiarisation Handler_mean | 0.31 | 0.27 | 1.36 | 0.77 | 2.35 |
| S Familiarisation Handler_prod | 0.04 | 0.46 | 1.04 | 0.93 | 1.16 |
| S Walking Pull_mean | -0.12 | 0.70 | 0.89 | 0.48 | 1.56 |
| S Walking Pull_prod | 0.00 | 0.96 | 1.00 | 0.87 | 1.14 |
| S Distractions Pull_mean ** | -0.58 | 0.05 | 0.56 | 0.30 | 0.99 |
| S Distractions Pull_prod | -0.10 | 0.06 | 0.91 | 0.82 | 1.00 |
| S Walking Distractions Pull_mean | -0.35 | 0.16 | 0.70 | 0.42 | 1.15 |
| S Walking Distractions Pull_prod ** | -0.09 | 0.06 | 0.91 | 0.83 | 1.00 |
| S Obedience_mean | 0.17 | 0.65 | 1.18 | 0.58 | 2.44 |
| S Obedience_prod | -0.00 | 0.97 | 1.00 | 0.97 | 1.03 |
| S Sensitivity_mean ** | -0.83 | 0.06 | 0.43 | 0.17 | 1.03 |
| S Tea Towel First Response | -0.30 | 0.24 | 0.74 | 0.44 | 1.21 |
| S Tea Towel Second Response | -0.08 | 0.72 | 0.92 | 0.59 | 1.43 |
| S Distractions First Response_mean | -0.26 | 0.43 | 0.77 | 0.39 | 1.45 |
| S Distractions First Response_prod | -0.00 | 0.30 | 1.00 | 0.99 | 1.00 |
| S Distractions Second Response_mean | -0.02 | 0.94 | 0.98 | 0.53 | 1.74 |
| S Distractions Second Response_prod | -0.00 | 0.37 | 1.00 | 0.99 | 1.00 |
| S Distractions Difference_mean | 0.29 | 0.40 | 1.34 | 0.67 | 2.67 |
| S Kong Response_mean | 0.34 | 0.43 | 1.41 | 0.59 | 3.31 |

Continued on next page.

Table C.7 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| S Kong Response_prod | 0.01 | 0.23 | 1.01 | 1.00 | 1.01 |
| S Dog Distraction_mean | -0.03 | 0.90 | 0.97 | 0.60 | 1.54 |
| S Dog Distraction_prod | -0.01 | 0.80 | 0.99 | 0.91 | 1.07 |
| S Call Response_mean | -0.05 | 0.91 | 0.96 | 0.41 | 2.10 |
| S Call Response_prod | -0.01 | 0.97 | 0.99 | 0.73 | 1.30 |
| S Familiarisation Dog Call | -0.05 | 0.87 | 0.95 | 0.55 | 1.70 |
| S Crate Response_mean | 0.01 | 0.97 | 1.01 | 0.48 | 2.12 |
| S Crate Response_prod | -0.01 | 0.22 | 0.99 | 0.96 | 1.01 |
| S Crate Stimulus_mean ** | 0.97 | 0.08 | 2.65 | 0.86 | 8.37 |
| S Crate Stimulus_prod | 0.01 | 0.38 | 1.01 | 0.98 | 1.05 |
| S Crate Handler_mean | -0.41 | 0.18 | 0.66 | 0.34 | 1.16 |
| S Crate Handler_prod | -0.03 | 0.19 | 0.97 | 0.91 | 1.01 |
| S Petting Stimulus During_mean *** | -1.31 | 0.02 | 0.27 | 0.08 | 0.75 |
| S Petting Stimulus During_prod ** | -0.39 | 0.06 | 0.68 | 0.43 | 0.95 |
| S Petting Engagement During_mean | -0.78 | 0.16 | 0.46 | 0.15 | 1.33 |
| S Petting Engagement During_prod | -0.11 | 0.51 | 0.89 | 0.61 | 1.21 |
| S Petting After_mean | -0.25 | 0.63 | 0.78 | 0.28 | 2.18 |
| S Petting After_prod | 0.09 | 0.51 | 1.09 | 0.82 | 1.41 |
| S Isolation Handler_mean | -0.08 | 0.84 | 0.93 | 0.43 | 1.95 |
| S Isolation Handler_prod | -0.00 | 0.98 | 1.00 | 0.96 | 1.04 |
| S Isolation Stimulus | -0.03 | 0.88 | 0.97 | 0.67 | 1.41 |
| S Sociability_mean ** | -0.76 | 0.06 | 0.47 | 0.21 | 1.01 |
| S Sociability_prod * | -0.02 | 0.14 | 0.98 | 0.96 | 1.00 |
| Items - Categorical | | | | | |
| Body check General Mouths | -15.32 | 0.99 | 0.00 | | |

Continued on next page.

Table C.7 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| Body check General Licks | -1.32 | 0.22 | 0.27 | 0.01 | 1.48 |
| Tea Towel First Response Indifferent * | 1.01 | 0.15 | 2.75 | 0.64 | 10.73 |
| Tea Towel First Response Change from Neutral | -0.56 | 0.40 | 0.57 | 0.16 | 2.31 |
| Tea Towel First Response Turns head | -0.13 | 0.84 | 0.88 | 0.27 | 3.42 |
| Tea Towel First Response Attempts to Removes towel by moving | -0.15 | 0.76 | 0.86 | 0.32 | 2.28 |
| Tea Towel First Response Attempts to Removes towel with mouth | -0.69 | 0.18 | 0.50 | 0.17 | 1.34 |
| Tea Towel First Response Plays | -0.99 | 0.36 | 0.37 | 0.02 | 2.16 |
| Tea Towel Second Response Indifferent | -0.02 | 0.98 | 0.98 | 0.29 | 2.93 |
| Tea Towel Second Response Turns head | 0.15 | 0.78 | 1.17 | 0.41 | 3.63 |
| Tea Towel Second Response Change from Neutral | 0.15 | 0.79 | 1.17 | 0.39 | 3.95 |
| Tea Towel Second Response Attempts to Removes towel by moving | -0.65 | 0.22 | 0.52 | 0.18 | 1.43 |
| Tea Towel Second Response Attempts to Removes towel with mouth | -0.21 | 0.68 | 0.81 | 0.28 | 2.21 |
| Tea Towel Second Response Plays | 0.14 | 0.85 | 1.15 | 0.24 | 4.25 |
| Crate Urinating | -14.18 | 0.99 | 0.00 | | |
| Sex | 0.62 | 0.22 | 1.85 | 0.70 | 5.16 |

95% Confidence Interval (CI): lower 2.5% and upper 97.5%
Statistically significant with $p < 0.05$ (***), $p < 0.10$ (**), $p < 0.15$ (*)

Table C.8: Univariate logistic regression model using the demographic variable Breed as a predictor of the probability of withdrawal, including estimates, p-value, Odds Ratio (OR) and Confidence Interval (CI).

| Variable | Estimate | P-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| Intercept | -0.00 | 1.00 | 1.00 | 0.04 | 25.29 |
| LR | -0.95 | 0.51 | 0.39 | 0.01 | 10.31 |
| LRxGR | -1.73 | 0.24 | 0.18 | 0.01 | 4.80 |

Table C.9: Univariate logistic regression model using the variable Kong Interaction Lateralisation as a predictor of the probability of withdrawal. Model parameters reported include estimates, p-value, Odds Ratio (OR) and Confidence Interval (CI).

| Variable | Estimate | P-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| Intercept | -1.39 | 0.08 | 0.25 | 0.04 | 1.00 |
| Left | 0.78 | 0.41 | 2.18 | 0.38 | 17.65 |
| None | -0.62 | 0.51 | 0.54 | 0.10 | 4.24 |
| Right | 0.41 | 0.66 | 1.50 | 0.27 | 11.84 |

### C.3.3 Visualisation

Figures C.1 to C.7 show the variables selected for the reduced dataset created with DC1 data. These variables were rated following the scales presented in Table C.1.

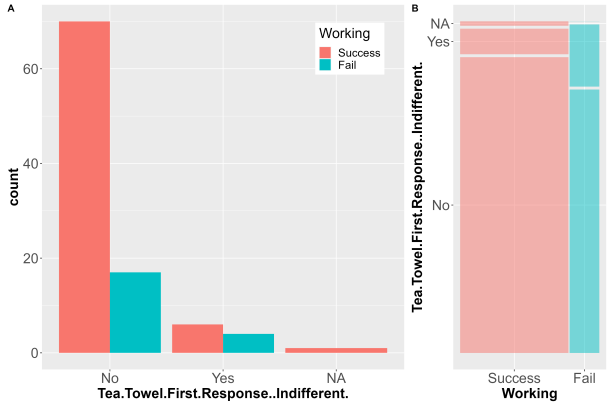



Figure C.1: Bar chart and box plot of the distribution of the ADTB item "Tea Towel First Response Indifferent" at DC1 in relation to training outcomes.

Figure C.2: Bar chart and box plot of the distribution of the ADTB item "Distractions First Response - Human" at DC1 in relation to training outcomes.

Figure C.3: Bar chart and box plot of the distribution of the ADTB item "Crate Behaviours - Actively Seeking Attention" at DC1 in relation to training outcomes.

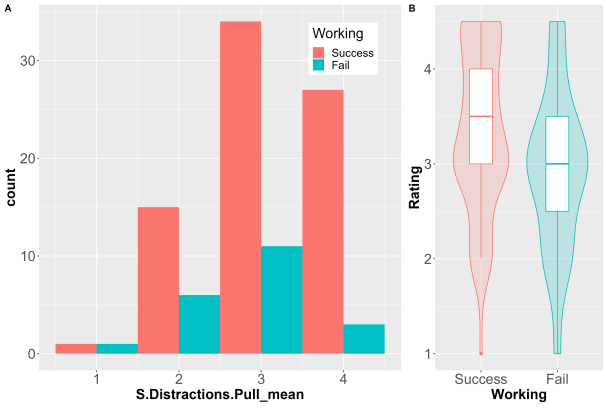



Figure C.4: Bar chart and box plot of the distribution of the ADTB score "S Distractions Pull_mean" at DC1 in relation to training outcomes.

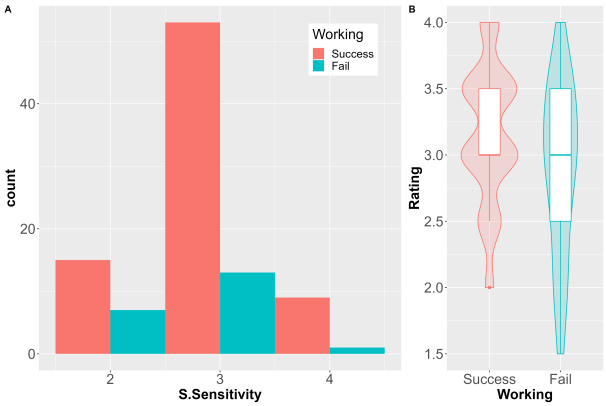



Figure C.5: Bar chart and box plot of the distribution of the ADTB score "S Sensitivity" at DC1 in relation to training outcomes.

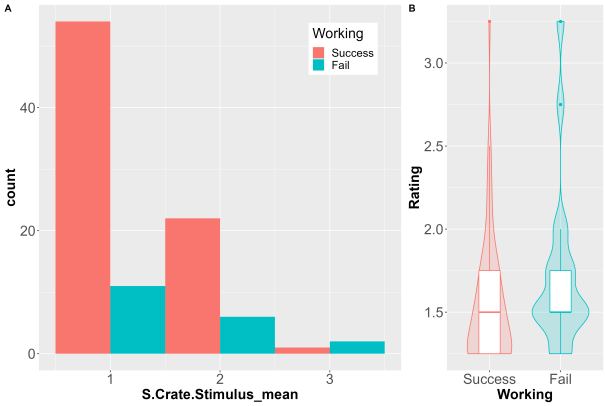



Figure C.6: Bar chart and box plot of the distribution of the ADTB score "S Crate Stimulus_mean" at DC1 in relation to training outcomes.

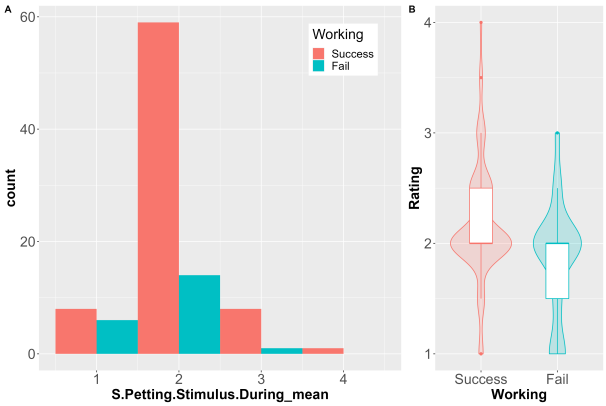



Figure C.7: Bar chart and box plot of the distribution of the ADTB score "S Petting Stimulus During_mean" at DC1 in relation to training outcomes.

# C.4 Data Collection 2 (DC2)

Table C.1 shows the ethogram used to mark the behaviours shown in each subtest.

## C.4.1 Descriptive Statistics

Table C.10 shows descriptive statistics of the ordinal ethogram items and composite scores of the data from DC2, including the minimum, 1st quartile, median, mean, 3rd quartile, maximum and number of missing values for all ethogram items and scores. Table C.11 shows descriptive statistics of the categorical variables including the number of dogs who scored “Yes”, “No” or are missing the score. The only categorical variable which has different categories is “Kong Interaction Lateralisation”, therefore the descriptive statistics are shown in Table C.12.

Table C.10: Descriptive statistics for DC2 numerical variables of the ADTB ethogram including minimum, 1st quartile, median, mean, 3rd quartile, maximum, and number of missing values for all items and scores.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| Familiarisation Response Oriented to Handler | 1.00 | 1.00 | 2.00 | 2.29 | 3.00 | 5.00 | 1 (0.02) |
| Familiarisation Response Exploration | 1.00 | 4.00 | 5.00 | 4.39 | 5.00 | 5.00 | 1 (0.02) |
| Familiarisation Response Waiting | 1.00 | 1.00 | 1.00 | 1.41 | 2.00 | 5.00 | 1 (0.02) |
| Call Back Response | 1.00 | 1.00 | 1.00 | 1.15 | 1.00 | 2.00 | 8 (0.13) |
| Walking Distractibility | 1.00 | 1.00 | 2.00 | 1.82 | 2.00 | 3.00 | 0 (0.00) |
| Walking Pull on leash | 1.00 | 1.00 | 2.00 | 1.77 | 2.00 | 4.00 | 0 (0.00) |
| Walking Pull strength | 1.00 | 1.00 | 1.50 | 1.53 | 2.00 | 3.00 | 0 (0.00) |
| Walking Initiative | 1.00 | 2.00 | 2.00 | 2.22 | 3.00 | 4.00 | 0 (0.00) |
| Standing Response | 1.00 | 2.00 | 3.00 | 2.80 | 4.00 | 4.00 | 0 (0.00) |
| Sitting Response | 1.00 | 2.00 | 3.00 | 2.88 | 4.00 | 4.00 | 0 (0.00) |
| Lying Response | 1.00 | 2.00 | 3.00 | 2.97 | 4.00 | 4.00 | 0 (0.00) |
| Lying Settled | 1.00 | 1.00 | 3.00 | 2.53 | 3.00 | 4.00 | 0 (0.00) |
| Body check Table | 1.00 | 3.00 | 4.00 | 3.40 | 4.00 | 4.00 | 0 (0.00) |
| Body check Response | 1.00 | 3.00 | 3.00 | 3.07 | 3.00 | 4.00 | 0 (0.00) |
| Distractions First Response Teddy | 1.00 | 2.00 | 3.00 | 3.02 | 4.00 | 6.00 | 0 (0.00) |
| Distractions First Response Human | 2.00 | 3.00 | 3.00 | 3.30 | 4.00 | 5.00 | 0 (0.00) |
| Distractions First Response Car | 1.00 | 2.00 | 3.00 | 2.97 | 4.00 | 6.00 | 0 (0.00) |
| Distractions First Response Food | 1.00 | 3.00 | 4.00 | 3.90 | 5.00 | 6.00 | 0 (0.00) |
| Distractions Second Response Teddy | 1.00 | 1.00 | 2.00 | 2.30 | 3.00 | 6.00 | 0 (0.00) |
| Distractions Second Response Human | 1.00 | 2.00 | 3.00 | 2.78 | 3.00 | 5.00 | 0 (0.00) |
| Distractions Second Response Car | 1.00 | 2.00 | 2.00 | 2.47 | 3.00 | 6.00 | 0 (0.00) |
| Distractions Second Response Food | 1.00 | 2.00 | 3.00 | 3.32 | 4.00 | 6.00 | 0 (0.00) |
| Distractions Pull on leash | 1.00 | 2.00 | 2.50 | 2.57 | 3.00 | 4.00 | 0 (0.00) |

Continued on next page.

Table C.10 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| Distractions Pull strength | 1.00 | 2.00 | 3.00 | 3.18 | 4.00 | 5.00 | 0 (0.00) |
| Distractions Human | 4.00 | 4.00 | 5.00 | 5.07 | 6.00 | 7.00 | 0 (0.00) |
| Kong Presentation Response | 2.00 | 2.00 | 4.00 | 3.58 | 4.25 | 5.00 | 0 (0.00) |
| Kong Interaction Response to stimulus | 2.00 | 3.00 | 3.00 | 3.07 | 3.00 | 4.00 | 1 (0.02) |
| Kong Interaction Response to handler | 1.00 | 1.00 | 1.00 | 1.65 | 2.00 | 3.00 | 0 (0.00) |
| Kong Interaction Back | 1.00 | 1.00 | 1.00 | 1.61 | 2.00 | 5.00 | 1 (0.02) |
| Kong Return Handler | 2.00 | 2.00 | 2.00 | 2.12 | 2.00 | 4.00 | 1 (0.02) |
| Dog Response | 1.00 | 4.00 | 5.00 | 4.87 | 6.00 | 8.00 | 0 (0.00) |
| Dog Call Back Response | 1.00 | 1.00 | 1.00 | 1.62 | 2.00 | 4.00 | 4 (0.07) |
| Crate Entering | 1.00 | 1.00 | 1.00 | 1.22 | 1.00 | 3.00 | 0 (0.00) |
| Crate Settled | 1.00 | 2.00 | 3.00 | 2.58 | 4.00 | 5.00 | 0 (0.00) |
| Crate Sniffing Exploring | 1.00 | 2.00 | 2.00 | 2.40 | 3.00 | 4.00 | 0 (0.00) |
| Crate Actively Seeking Attention | 1.00 | 1.00 | 1.00 | 1.83 | 3.00 | 5.00 | 0 (0.00) |
| Crate Digging | 1.00 | 1.00 | 1.00 | 1.15 | 1.00 | 3.00 | 0 (0.00) |
| Crate Nudging Crate | 1.00 | 1.00 | 1.00 | 1.22 | 1.00 | 3.00 | 0 (0.00) |
| Crate Pawing | 1.00 | 1.00 | 1.00 | 1.22 | 1.00 | 3.00 | 0 (0.00) |
| Crate Whining | 1.00 | 1.00 | 1.00 | 1.68 | 2.00 | 5.00 | 3 (0.05) |
| Crate Barking | 1.00 | 1.00 | 1.00 | 1.15 | 1.00 | 5.00 | 0 (0.00) |
| Crate Self Modulation | 1.00 | 1.00 | 2.00 | 1.72 | 2.00 | 3.00 | 0 (0.00) |
| Crate Stress | 1.00 | 1.75 | 2.00 | 1.97 | 2.00 | 4.00 | 0 (0.00) |
| Petting Handler Stimulus | 1.00 | 1.00 | 1.00 | 1.14 | 1.00 | 3.00 | 1 (0.02) |
| Petting Handler Holding dog | 1.00 | 1.00 | 1.00 | 1.10 | 1.00 | 3.00 | 1 (0.02) |
| Petting Confidence During | 1.00 | 3.00 | 3.00 | 3.19 | 4.00 | 5.00 | 1 (0.02) |
| Petting Responsiveness After | 1.00 | 3.00 | 4.00 | 3.51 | 4.00 | 5.00 | 1 (0.02) |
| Isolation Response Time oriented | 1.00 | 3.00 | 4.00 | 4.00 | 5.00 | 5.00 | 0 (0.00) |

Continued on next page.

Table C.10 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| Isolation Response Exploration | 1.00 | 1.00 | 2.00 | 2.02 | 3.00 | 5.00 | 0 (0.00) |
| Isolation Response Unsettled Pacing | 1.00 | 1.00 | 1.50 | 1.70 | 2.00 | 4.00 | 0 (0.00) |
| Isolation Response Whining | 1.00 | 1.00 | 1.00 | 1.69 | 2.00 | 4.00 | 8 (0.13) |
| Isolation Response Barking | 1.00 | 1.00 | 1.00 | 1.29 | 1.00 | 4.00 | 1 (0.02) |
| Isolation Reunion | 1.00 | 4.00 | 4.00 | 4.08 | 4.00 | 6.00 | 0 (0.00) |
| Noise Confidence | 3.00 | 4.00 | 4.00 | 3.97 | 4.00 | 6.00 | 0 (0.00) |
| S Familiarisation Handler_mean | 1.00 | 1.00 | 1.50 | 1.85 | 2.00 | 5.00 | 1 (0.02) |
| S Familiarisation Handler_prod | 1.00 | 1.00 | 2.00 | 3.72 | 4.00 | 25.00 | 0 (0.00) |
| S Walking Pull_mean | 1.00 | 1.00 | 1.75 | 1.65 | 2.00 | 3.50 | 0 (0.00) |
| S Walking Pull_prod | 1.00 | 1.00 | 3.00 | 3.12 | 4.00 | 12.00 | 0 (0.00) |
| S Distractions Pull_mean | 1.00 | 2.00 | 3.00 | 2.88 | 3.50 | 4.50 | 0 (0.00) |
| S Distractions Pull_prod | 1.00 | 4.00 | 8.00 | 8.78 | 12.00 | 20.00 | 0 (0.00) |
| S Walking Distractions Pull_mean | -0.50 | 0.88 | 1.00 | 1.23 | 2.00 | 3.00 | 0 (0.00) |
| S Walking Distractions Pull_prod | -2.00 | 2.75 | 5.00 | 5.67 | 8.00 | 18.00 | 0 (0.00) |
| S Obedience_mean | 1.33 | 2.33 | 3.00 | 2.88 | 3.33 | 4.00 | 0 (0.00) |
| S Obedience_prod | 2.00 | 9.00 | 21.00 | 24.37 | 36.00 | 64.00 | 0 (0.00) |
| S Sensitivity_mean | 1.50 | 3.00 | 3.50 | 3.23 | 3.50 | 4.00 | 0 (0.00) |
| S Tea Towel First Response | 0.00 | 1.00 | 1.67 | 1.64 | 2.50 | 3.50 | 0 (0.00) |
| S Tea Towel Second Response | 0.00 | 0.00 | 1.00 | 1.21 | 2.00 | 3.50 | 0 (0.00) |
| S Distractions First Response_mean | 1.75 | 2.75 | 3.25 | 3.30 | 3.75 | 5.75 | 0 (0.00) |
| S Distractions First Response_prod | 6.00 | 48.00 | 93.00 | 141.05 | 153.00 | 1080.00 | 0 (0.00) |
| S Distractions Second Response_mean | 1.25 | 2.25 | 2.75 | 2.72 | 3.00 | 5.75 | 0 (0.00) |
| S Distractions Second Response_prod | 2.00 | 16.00 | 44.00 | 79.72 | 72.00 | 1080.00 | 0 (0.00) |
| S Distractions Difference_mean | -2.50 | -1.00 | -0.50 | -0.58 | 0.00 | 0.75 | 0 (0.00) |
| S Kong Response_mean | 1.75 | 2.25 | 2.50 | 2.59 | 2.75 | 4.50 | 0 (0.00) |

Continued on next page.

Table C.10 - continued from previous page.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs (%) |
|---|---|---|---|---|---|---|---|
| S Kong Response_prod | 2.00 | 24.00 | 24.00 | 43.87 | 48.00 | 400.00 | 0 (0.00) |
| S Dog Distraction_mean | 1.00 | 2.50 | 3.00 | 3.37 | 4.00 | 6.00 | 0 (0.00) |
| S Dog Distraction_prod | 1.00 | 4.00 | 5.00 | 8.23 | 12.00 | 28.00 | 0 (0.00) |
| S Call Response_mean | 1.00 | 1.00 | 1.00 | 1.38 | 1.50 | 3.00 | 1 (0.02) |
| S Call Response_prod | 1.00 | 1.00 | 1.00 | 1.83 | 2.00 | 8.00 | 0 (0.00) |
| S Familiarisation Dog Call | -3.00 | -1.00 | 0.00 | -0.51 | 0.00 | 1.00 | 11 (0.18) |
| S Crate Response_mean | 0.75 | 1.25 | 1.75 | 1.83 | 2.25 | 3.25 | 0 (0.00) |
| S Crate Response_prod | 0.00 | 2.00 | 8.00 | 15.32 | 16.00 | 108.00 | 0 (0.00) |
| S Crate Stimulus_mean | 1.00 | 1.25 | 1.50 | 1.50 | 1.75 | 3.00 | 0 (0.00) |
| S Crate Stimulus_prod | 1.00 | 2.00 | 3.00 | 5.27 | 4.00 | 72.00 | 0 (0.00) |
| S Crate Handler_mean | 1.00 | 1.00 | 1.00 | 1.54 | 2.00 | 3.67 | 0 (0.00) |
| S Crate Handler_prod | 1.00 | 1.00 | 1.00 | 5.42 | 6.00 | 45.00 | 0 (0.00) |
| S Petting Stimulus During_mean | 1.00 | 2.00 | 2.00 | 2.16 | 2.50 | 4.00 | 1 (0.02) |
| S Petting Stimulus During_prod | 1.00 | 3.00 | 3.00 | 3.73 | 4.00 | 15.00 | 0 (0.00) |
| S Petting Engagement During_mean | 1.00 | 2.00 | 2.00 | 2.14 | 2.50 | 3.00 | 1 (0.02) |
| S Petting Engagement During_prod | 1.00 | 3.00 | 3.00 | 3.42 | 4.00 | 9.00 | 0 (0.00) |
| S Petting After_mean | 1.00 | 2.00 | 2.50 | 2.31 | 2.50 | 3.50 | 1 (0.02) |
| S Petting After_prod | 1.00 | 3.00 | 4.00 | 3.80 | 4.00 | 12.00 | 0 (0.00) |
| S Isolation Handler_mean | 1.00 | 2.00 | 2.33 | 2.39 | 2.67 | 4.00 | 0 (0.00) |
| S Isolation Handler_prod | 1.00 | 4.00 | 5.00 | 9.95 | 8.50 | 64.00 | 0 (0.00) |
| S Isolation Stimulus | -2.00 | -0.25 | 0.00 | 0.32 | 1.00 | 4.00 | 0 (0.00) |
| S Sociability_mean | 1.00 | 3.33 | 3.67 | 3.61 | 4.00 | 5.00 | 0 (0.00) |
| S Sociability_prod | 1.00 | 32.00 | 48.00 | 48.03 | 60.00 | 120.00 | 0 (0.00) |

Table C.11: Descriptive statistics of DC2 for categorical variables of the ADTB ethogram including the number of No, Yes and missing data.

| Variable | No | Yes | NAs (%) |
|---|---|---|---|
| Body check General Mouths | 60 | 60 | 0 (0.00) |
| Body check General Licks | 56 | 4 | 0 (0.00) |
| Tea Towel First Response Indifferent | 52 | 8 | 0 (0.00) |
| Tea Towel First Response Change from Neutral | 9 | 51 | 0 (0.00) |
| Tea Towel First Response Turns head | 12 | 48 | 0 (0.00) |
| Tea Towel First Response Attempts to Removes towel by moving | 33 | 27 | 0 (0.00) |
| Tea Towel First Response Attempts to Removes towel with mouth | 33 | 27 | 0 (0.00) |
| Tea Towel First Response Plays | 52 | 8 | 0 (0.00) |
| Tea Towel Second Response Indifferent | 38 | 22 | 0 (0.00) |
| Tea Towel Second Response Turns head | 28 | 32 | 0 (0.00) |
| Tea Towel Second Response Change from Neutral | 21 | 39 | 0 (0.00) |
| Tea Towel Second Response Attempts to Removes towel by moving | 40 | 20 | 0 (0.00) |
| Tea Towel Second Response Attempts to Removes towel with mouth | 40 | 20 | 0 (0.00) |
| Tea Towel Second Response Plays | 53 | 7 | 0 (0.00) |
| Crate Urinating | 60 | 60 | 0 (0.00) |
| Isolation Urinating | 59 | 1 | 0 (0.00) |

Table C.12: ADTB ethogram descriptive statistics for the variable Kong Interaction Lateralisation, including the number of dogs in the sample per category.

| Variable | None | Both | Right | Left |
|---|---|---|---|---|
| Kong - Interaction Lateralisation | 20 | 12 | 11 | 16 |

Table C.13: Descriptive statistics of the features used in the optimal multivariate logistic regression model and the items from which they derive.

| Variable | Min | $Q_1$ | Median | Mean | $Q_3$ | Max | NAs |
|---|---|---|---|---|---|---|---|
| Distractions First Response Car | 1.00 | 2.00 | 3.00 | 2.97 | 4.00 | 6.00 | 0.00 |
| S Familiarisation Handler_mean | 1.00 | 1.00 | 1.50 | 1.85 | 2.00 | 5.00 | 1 |
| Familiarisation Response Oriented to Handler | 1.00 | 1.00 | 2.00 | 2.29 | 3.00 | 5.00 | 1 |
| Familiarisation Response Waiting | 1.00 | 1.00 | 1.00 | 1.41 | 2.00 | 5.00 | 1 |
| S Walking Pull_mean | 1.00 | 1.00 | 1.75 | 1.65 | 2.00 | 3.50 | 0 |
| Walking Pull on leash | 1.00 | 1.00 | 2.00 | 1.77 | 2.00 | 4.00 | 0 |
| Walking Pull strength | 1.00 | 1.00 | 1.50 | 1.53 | 2.00 | 3.00 | 0 |
| S Sensitivity_mean | 1.50 | 3.00 | 3.50 | 3.23 | 3.50 | 4.00 | 0 |
| Body check Table | 1.00 | 3.00 | 4.00 | 3.40 | 4.00 | 4.00 | 0 |
| Body check Response | 1.00 | 3.00 | 3.00 | 3.07 | 3.00 | 4.00 | 0 |
| S Crate Handler_mean | 1.00 | 1.00 | 1.00 | 1.54 | 2.00 | 3.67 | 0 |
| Crate Actively Seeking Attention | 1.00 | 1.00 | 1.00 | 1.83 | 3.00 | 5.00 | 0 |
| Crate Whining | 1.00 | 1.00 | 1.00 | 1.68 | 2.00 | 5.00 | 3 |
| Crate Barking | 1.00 | 1.00 | 1.00 | 1.15 | 1.00 | 5.00 | 0 |
| S Isolation Stimulus | -2.00 | -0.25 | 0.00 | 0.32 | 1.00 | 4.00 | 0 |
| Isolation Response Exploration | 1.00 | 1.00 | 2.00 | 2.02 | 3.00 | 5.00 | 0 |
| Isolation Response Unsettled/Pacing | 1.00 | 1.00 | 1.50 | 1.70 | 2.00 | 4.00 | 0 |

## C.4.2 Univariate Analysis

Multiple models were built using each ethogram variable individually to predict the probability of withdrawal from training. The results from

the univariate logistic regression models are shown in C.14. It includes the model estimate, p-value, Odds Ratio, and the lower 2.5% and upper 97.5% bounds of the Confidence Interval (CI). Positive estimates indicate that the variable was positively associated with a higher probability of withdrawal. Negative estimates suggest that the variable was negatively associated with a higher probability of withdrawal. The p-value shows the level of significance of this relationship.

Table C.14: ADTB univariate logistic regression model estimates, p-value, Odds Ratio (OR), and Confidence Interval (CI).

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| Familiarisation Response Oriented to Handler | -0.48 | 0.15 | 0.62 | 0.29 | 1.12 |
| Familiarisation Response Exploration | 0.15 | 0.67 | 1.16 | 0.63 | 2.69 |
| Familiarisation Response Waiting | -1.43 | 0.17 | 0.24 | 0.01 | 1.01 |
| Call Back Response | -0.85 | 0.45 | 0.43 | 0.02 | 2.81 |
| Walking Distractibility | 0.03 | 0.93 | 1.03 | 0.45 | 2.31 |
| Walking Pull on leash *** | -1.44 | 0.03 | 0.24 | 0.05 | 0.71 |
| Walking Pull strength *** | -1.90 | 0.02 | 0.15 | 0.02 | 0.62 |
| Walking Initiative | 0.42 | 0.31 | 1.53 | 0.67 | 3.54 |
| Standing Response | 0.50 | 0.16 | 1.65 | 0.85 | 3.55 |
| Sitting Response | -0.14 | 0.63 | 0.87 | 0.48 | 1.57 |
| Lying Response | -0.25 | 0.34 | 0.78 | 0.47 | 1.31 |
| Lying Settled | -0.04 | 0.91 | 0.97 | 0.54 | 1.76 |
| Body check Table * | -0.42 | 0.14 | 0.66 | 0.37 | 1.17 |
| Body check Response | -0.45 | 0.36 | 0.64 | 0.23 | 1.73 |
| Distractions First Response Teddy | -0.14 | 0.59 | 0.87 | 0.51 | 1.41 |
| Distractions First Response Human | 0.19 | 0.61 | 1.21 | 0.59 | 2.56 |
| Distractions First Response Car *** | 0.79 | 0.02 | 2.21 | 1.18 | 4.71 |
| Distractions First Response Food | 0.18 | 0.46 | 1.19 | 0.76 | 1.98 |
| Distractions Second Response Teddy | -0.04 | 0.87 | 0.96 | 0.53 | 1.59 |
| Distractions Second Response Human | -0.19 | 0.61 | 0.82 | 0.38 | 1.70 |
| Distractions Second Response Car | 0.39 | 0.19 | 1.47 | 0.82 | 2.69 |
| Distractions Second Response Food | 0.21 | 0.40 | 1.24 | 0.74 | 2.08 |
| Distractions Pull on leash | 0.04 | 0.93 | 1.04 | 0.44 | 2.39 |
| Distractions Human | 0.03 | 0.94 | 1.03 | 0.48 | 2.19 |

Continued on next page.

Table C.14 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| Kong Presentation Response | -0.15 | 0.59 | 0.86 | 0.51 | 1.48 |
| Kong Interaction Response to stimulus *** | 2.16 | 0.03 | 8.63 | 1.31 | 73.31 |
| Kong Interaction Response to handler | 0.03 | 0.94 | 1.03 | 0.46 | 2.15 |
| Kong Interaction Back | 0.37 | 0.25 | 1.44 | 0.74 | 2.74 |
| Kong Return Handler ** | 1.27 | 0.06 | 3.58 | 0.96 | 17.30 |
| Dog Response | 0.03 | 0.89 | 1.03 | 0.65 | 1.65 |
| Dog Call Back Response | 0.35 | 0.34 | 1.42 | 0.67 | 2.93 |
| Crate Entering | -0.26 | 0.71 | 0.77 | 0.13 | 2.51 |
| Crate Settled | -0.00 | 1.00 | 1.00 | 0.57 | 1.77 |
| Crate Sniffing Exploring | 0.28 | 0.43 | 1.33 | 0.65 | 2.73 |
| Crate Actively Seeking Attention ** | -0.87 | 0.08 | 0.42 | 0.12 | 0.92 |
| Crate Digging | 0.54 | 0.39 | 1.71 | 0.42 | 5.91 |
| Crate Whining | -0.81 | 0.12 | 0.45 | 0.12 | 1.03 |
| Crate Barking | -15.72 | 0.99 | 0.00 | | |
| Crate Self Modulation | -0.12 | 0.79 | 0.89 | 0.35 | 2.09 |
| Crate Nudging Crate | -1.02 | 0.31 | 0.36 | 0.02 | 1.71 |
| Crate Pawing | -15.61 | 0.99 | 0.00 | | |
| Crate Stress | -0.13 | 0.78 | 0.88 | 0.34 | 2.15 |
| Petting Handler Stimulus | -0.53 | 0.61 | 0.59 | 0.03 | 3.04 |
| Petting Handler Holding dog | -0.16 | 0.86 | 0.86 | 0.06 | 3.62 |
| Petting Confidence During | -0.34 | 0.39 | 0.71 | 0.31 | 1.53 |
| Petting Responsiveness After | 0.38 | 0.31 | 1.47 | 0.72 | 3.19 |
| Isolation Response Time oriented *** | -0.71 | 0.03 | 0.49 | 0.24 | 0.92 |
| Isolation Response Exploration *** | 0.60 | 0.05 | 1.83 | 1.02 | 3.45 |
| Isolation Response Unsettled Pacing | -0.40 | 0.36 | 0.67 | 0.25 | 1.47 |

Continued on next page.

Table C.14 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| Isolation Response Whining | -0.41 | 0.35 | 0.66 | 0.24 | 1.42 |
| Isolation Response Barking | -0.65 | 0.37 | 0.52 | 0.04 | 1.41 |
| Isolation Reunion | -0.28 | 0.45 | 0.76 | 0.37 | 1.60 |
| Noise Confidence | -0.32 | 0.49 | 0.73 | 0.27 | 1.70 |
| S Familiarisation Handler_mean | -0.96 | 0.10 | 0.38 | 0.10 | 1.01 |
| S Familiarisation Handler_prod | -0.35 | 0.15 | 0.70 | 0.39 | 0.99 |
| S Walking Pull_mean *** | -1.72 | 0.02 | 0.18 | 0.03 | 0.63 |
| S Walking Pull_prod *** | -0.57 | 0.02 | 0.56 | 0.31 | 0.86 |
| S Distractions Pull_mean | -0.07 | 0.86 | 0.94 | 0.45 | 1.93 |
| S Distractions Pull_prod | 0.01 | 0.92 | 1.01 | 0.88 | 1.14 |
| S Walking Distractions Pull_mean ** | 0.78 | 0.06 | 2.19 | 0.98 | 5.34 |
| S Walking Distractions Pull_prod | 0.10 | 0.16 | 1.10 | 0.96 | 1.27 |
| S Obedience_mean | -0.06 | 0.90 | 0.94 | 0.36 | 2.54 |
| S Obedience_prod | -0.00 | 0.99 | 1.00 | 0.96 | 1.04 |
| S Sensitivity_mean ** | -0.86 | 0.09 | 0.42 | 0.15 | 1.14 |
| S Tea Towel First Response | -0.05 | 0.88 | 0.95 | 0.49 | 1.81 |
| S Tea Towel Second Response | 0.03 | 0.92 | 1.03 | 0.57 | 1.79 |
| S Distractions First Response_mean | 0.46 | 0.26 | 1.59 | 0.70 | 3.64 |
| S Distractions First Response_prod | 0.00 | 0.81 | 1.00 | 1.00 | 1.00 |
| S Distractions Second Response_mean | 0.22 | 0.57 | 1.24 | 0.55 | 2.65 |
| S Distractions Second Response_prod | -0.00 | 0.99 | 1.00 | 0.99 | 1.00 |
| S Distractions Difference_mean | -0.27 | 0.55 | 0.76 | 0.31 | 1.90 |
| S Kong Response_mean | 0.53 | 0.38 | 1.70 | 0.49 | 5.66 |
| S Kong Response_prod | 0.01 | 0.15 | 1.01 | 1.00 | 1.02 |
| S Dog Distraction_mean | 0.01 | 0.98 | 1.01 | 0.55 | 1.78 |

Continued on next page.

Table C.14 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| S Dog Distraction_prod | 0.03 | 0.61 | 1.03 | 0.92 | 1.14 |
| S Call Response_mean | 0.39 | 0.54 | 1.48 | 0.39 | 5.06 |
| S Call Response_prod | 0.06 | 0.81 | 1.06 | 0.61 | 1.61 |
| S Familiarisation Dog Call | -0.38 | 0.30 | 0.68 | 0.32 | 1.40 |
| S Crate Response_mean | -0.12 | 0.82 | 0.88 | 0.30 | 2.44 |
| S Crate Response_prod | -0.02 | 0.40 | 0.98 | 0.92 | 1.01 |
| S Crate Stimulus_mean | -0.15 | 0.87 | 0.86 | 0.13 | 4.26 |
| S Crate Stimulus_prod | -0.03 | 0.61 | 0.97 | 0.79 | 1.04 |
| S Crate Handler_mean * | -1.25 | 0.09 | 0.29 | 0.04 | 0.92 |
| S Crate Handler_prod | -0.22 | 0.14 | 0.80 | 0.53 | 0.98 |
| S Petting Stimulus During_mean | -0.57 | 0.39 | 0.57 | 0.13 | 1.90 |
| S Petting Stimulus During_prod | -0.15 | 0.43 | 0.86 | 0.52 | 1.16 |
| S Petting Engagement During_mean | -0.69 | 0.36 | 0.50 | 0.11 | 2.19 |
| S Petting Engagement During_prod | -0.14 | 0.62 | 0.87 | 0.48 | 1.42 |
| S Petting After_mean | 0.60 | 0.38 | 1.82 | 0.49 | 7.33 |
| S Petting After_prod | 0.11 | 0.52 | 1.12 | 0.76 | 1.58 |
| S Isolation Handler_mean *** | -1.23 | 0.04 | 0.29 | 0.08 | 0.87 |
| S Isolation Handler_prod | -0.04 | 0.35 | 0.96 | 0.86 | 1.02 |
| S Isolation Stimulus *** | 0.60 | 0.03 | 1.83 | 1.10 | 3.25 |
| S Sociability_mean | -0.14 | 0.76 | 0.87 | 0.35 | 2.20 |
| S Sociability_prod | -0.01 | 0.61 | 0.99 | 0.97 | 1.02 |
| Body check General Licks | 0.31 | 0.80 | 1.36 | 0.06 | 11.88 |
| Tea Towel First Response Indifferent | -17.36 | 0.99 | 0.00 | | |
| Tea Towel First Response Change from Neutral | 17.39 | 0.99 | 35584245.59 | 0.00 | |
| Tea Towel First Response Turns head | 1.18 | 0.28 | 3.27 | 0.54 | 63.12 |

Continued on next page.

Table C.14 - continued from previous page.

| Variable | Estimate | p-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| Tea Towel First Response Attempts to Removes towel by moving | -0.17 | 0.80 | 0.84 | 0.22 | 3.02 |
| Tea Towel First Response Attempts to Removes towel with mouth | -0.61 | 0.37 | 0.54 | 0.13 | 1.97 |
| Tea Towel First Response Plays | -0.63 | 0.57 | 0.53 | 0.03 | 3.47 |
| Tea Towel Second Response Indifferent *** | -2.15 | 0.05 | 0.12 | 0.01 | 0.67 |
| Tea Towel Second Response Turns head | 0.69 | 0.31 | 2.00 | 0.55 | 8.33 |
| Tea Towel Second Response Change from Neutral ** | 2.06 | 0.06 | 7.86 | 1.36 | 149.47 |
| Tea Towel Second Response Attempts to Removes towel by moving | -0.00 | 1.00 | 1.00 | 0.24 | 3.70 |
| Tea Towel Second Response Attempts to Removes towel with mouth | -1.10 | 0.19 | 0.33 | 0.05 | 1.45 |
| Tea Towel Second Response Plays | -0.45 | 0.69 | 0.64 | 0.03 | 4.29 |
| Sex | 1.03 | 0.13 | 2.80 | 0.77 | 11.70 |

95% Confidence Interval (CI): lower 2.5% and upper 97.5%
Statistically significant with $p < 0.05$ (***), $p < 0.10$ (**), $p < 0.15$ (*)

Table C.15: Univariate logistic regression model using the demographic variable Breed as a predictor of the probability of withdrawal, including estimates, p-value, Odds Ratio (OR) and Confidence Interval (CI).

| Variable | Estimate | P-value |
|---|---|---|
| (Intercept) | 16.57 | 0.99 |
| LR | -17.95 | 0.99 |
| LRxGR | -18.11 | 0.99 |

Table C.16: Univariate logistic regression model using the variable Kong Interaction Lateralisation as a predictor of the probability of withdrawal. Model parameters reported include estimates, p-value, Odds Ratio (OR) and Confidence Interval (CI).

| Variable | Estimate | P-value | OR | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| (Intercept) | 0.34 | 0.57 | 1.40 | 0.45 | 4.73 |
| Left | 1.20 | 0.16 | 3.33 | 0.63 | 20.46 |
| None | 1.86 | 0.05 | 6.43 | 1.11 | 52.98 |
| Right | 1.97 | 0.10 | 7.14 | 0.89 | 153.54 |

## C.4.3 Visualisation

Figures C.8 to C.16 show the variables selected for the reduced dataset created with DC2 data. These variables were rated following the scales presented in Table C.1.

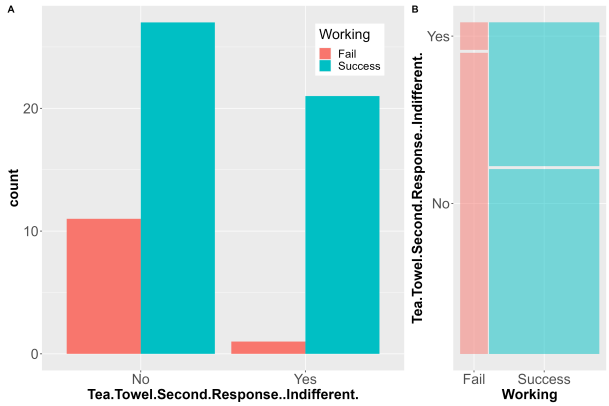



Figure C.8: Bar chart and box plot of the distribution of the ADTB item "Tea Towel Second Response Indifferent" at DC2 in relation to training outcomes.

Figure C.9: Bar chart and box plot of the distribution of the ADTB item "Distractions First Response Car" at DC2 in relation to training outcomes.

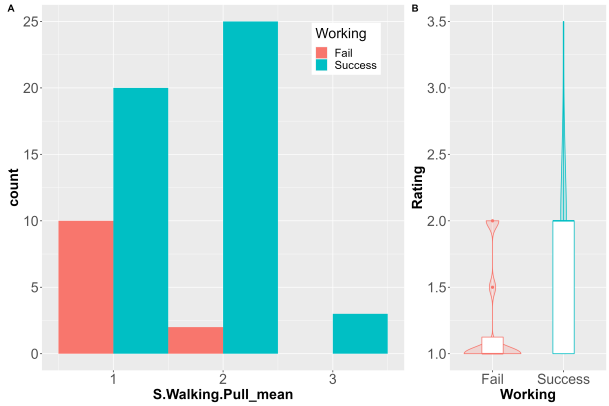



Figure C.10: Bar chart and box plot of the distribution of the ADTB score "S Walking Pull_mean" at DC2 in relation to training outcomes.

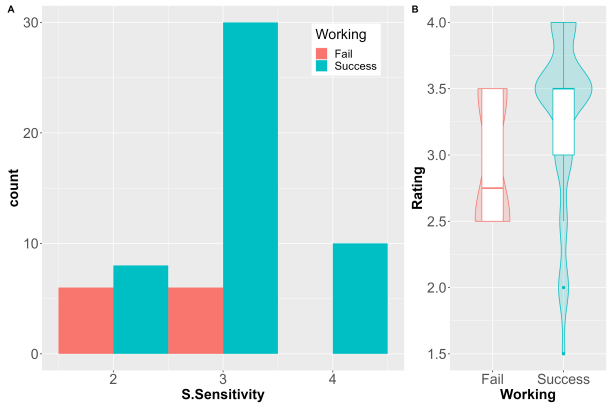



Figure C.11: Bar chart and box plot of the distribution of the ADTB score "S Sensitivity" at DC2 in relation to training outcomes.

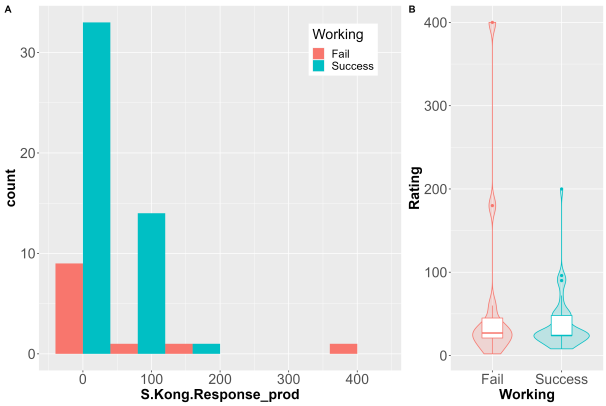



Figure C.12: Bar chart and box plot of the distribution of the ADTB score "S Kong Response_prod" at DC2 in relation to training outcomes.

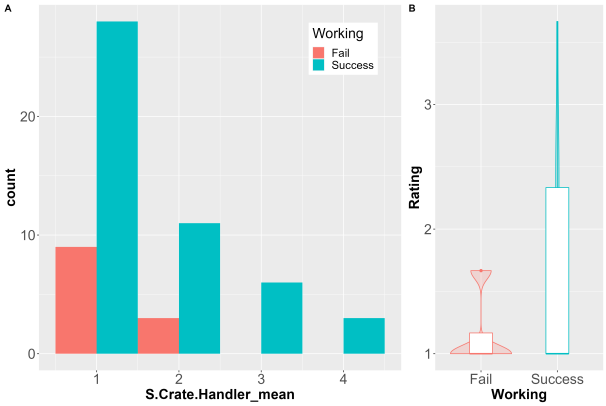



Figure C.13: Bar chart and box plot of the distribution of the ADTB score "S Crate Handler_mean" at DC2 in relation to training outcomes.

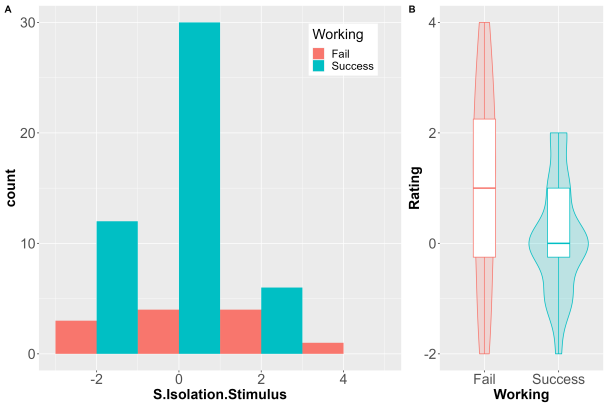



Figure C.14: Bar chart and box plot of the distribution of the ADTB score "S Isolation Stimulus" at DC2 in relation to training outcomes.

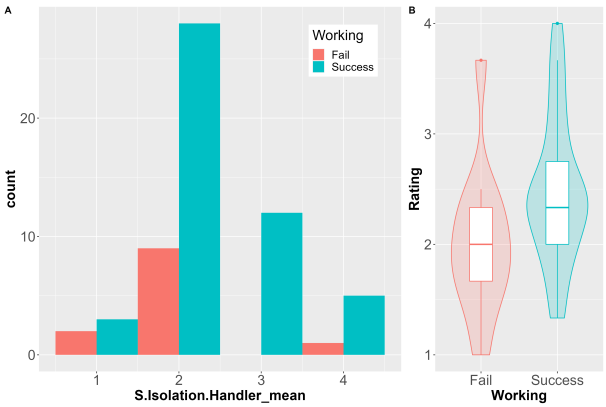



Figure C.15: Bar chart and box plot of the distribution of the ADTB score "S Isolation Handler_mean" at DC2 in relation to training outcomes.

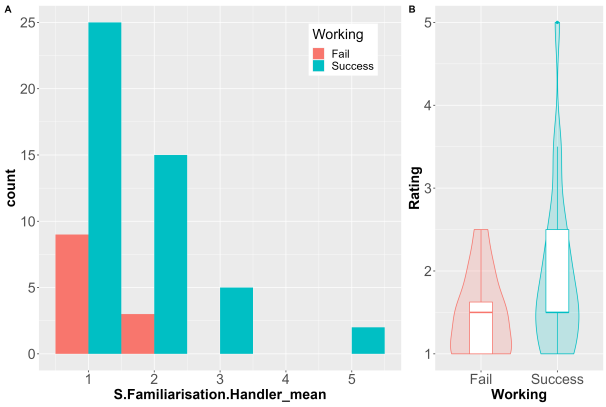



Figure C.16: Bar chart and box plot of the distribution of the ADTB score "S Familiarisation Handler_mean" at DC2 in relation to training outcomes.