CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Prediction of working outcomes in trainee dogs using the novel Assistance Dog Test Battery (ADTB)
Authors	Marcato, Marinara;Tedesco, Salvatore;O'Mahony, Conor;O'Flynn, Brendan;Galvin, Paul
Publication date	2024-02-28
Original Citation	Marcato, M., Tedesco, S., O'Mahony, C., O'Flynn, B. and Galvin, P. (2024) 'Prediction of working outcomes in trainee dogs using the novel Assistance Dog Test Battery (Adtb)', Applied Animal Behaviour Science, 272, 106212 (11 pp.). Available at: https://doi.org/10.1016/j.applanim.2024.106212
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://doi.org/10.17632/5mzfpt455r.1 - 10.1016/j.applanim.2024.106212https://github.com/mmarcato/dog_ethogram/
Rights	© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/) - http://creativecommons.org/licenses/by/4.0/
Download date	2026-11-10 17:16:23
Item downloaded from	https://hdl.handle.net/10468/15665



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

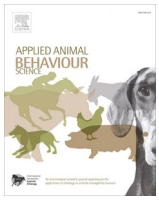
Applied Animal Behaviour Science 272 (2024) 106212



Contents lists available at ScienceDirect

Applied Animal Behaviour Science

journal homepage: www.elsevier.com/locate/applanim



Prediction of working outcomes in trainee dogs using the novel Assistance Dog Test Battery (ADTB)

Marinara Marcato [a,*,1], Salvatore Tedesco [a], Conor O'Mahony [a], Brendan O'Flynn [a], Paul Galvin [a]

[a] Tyndall National Institute, Lee Maltings Complex Dyke Parade, Cork T12 R5CP, Ireland

ARTICLE INFO

Keywords:
Assistance dogs
Guide dogs
Service dogs
Outcome prediction
ADTB
Canine behaviour tests
Dog behaviour assessment

ABSTRACT

Canine behaviour is commonly assessed using test batteries comprising a test protocol and ethogram scoring system. These are particularly valuable for assistance dog organisations as a tool for evaluating trainee dogs' proficiency in fundamental skills. The goal of this study was to design and validate a new test battery to assess the suitability of trainee dogs for assistance work at different stages of the training programme. The main objective was to develop a machine-learning tool capable of predicting working outcomes. Accordingly, the novel Assistance Dog Test Battery (ADTB) was developed. Trainee assistance dogs participating in this research performed the test at 3 weeks and 10 weeks after starting formal training. The results from the univariate logistic regression analysis were used to select the variables for the reduced feature sets that were used for modelling. The machine learning models were built using the data collected at 3 and 10 weeks separately and predicted working outcomes with an area under the ROC curve of 0.74 and 0.84, respectively. This research demonstrated the relationship between the novel ADTB ethogram measures and working outcomes in assistance dogs. The machine learning model created using the data collected at 3 weeks achieved comparable performance to the state-of-the-art, while the model built using the data collected at 10 weeks substantially outperformed it. These preliminary results suggest that the ADTB is a reliable tool for the prediction of working outcomes in trainee assistance dogs. Hence, assistance dog organisations can reduce the cost of training by using model predictions as a guide for deciding which dogs to withdraw from training. The data collected and the code developed in this research are publicly available on Mendeley Data (https://doi.org/10.17632/5mzfpt455r.1) and GitHub, respectively (https://github.com/mmarcato/dog_ethogram/).

1. Introduction

Working dogs are used in specific roles for economic reasons and their natural aptitude for the task at hand. Assistance dogs include service dogs, which provide support to individuals with physical, medical, or emotional disabilities; and guide dogs, which assist individuals with visual impairments (ADI, 2022. Training failure rates of 50–70% are typical and the production costs reach €53,000 in Ireland (IGDB, 2021. Accurately evaluating proficiency in fundamental skills and predicting successful qualification as early as possible in the training program is crucial to increasing efficiency and reducing costs (Bray et al., 2019.

The most commonly used canine behaviour assessment method are test batteries - also known as behaviour tests (Jones and Gosling, 2005 or behavioural assessments (Bray et al., 2021. Test batteries take place in a controlled environment where stimuli are systematically presented to an animal according to a subtest protocol and their behaviour is assessed using an ethogram marking system. Six studies were found to develop predictive models of training outcomes in assistance dogs using different test batteries (Batt et al., 2008; Asher et al., 2013; Harvey et al., 2016; Bray et al., 2017b, 2019; MacLean and Hare, 2018). Each employed different outcome prediction techniques, performance estimation methods, number of subjects, age at assessment and class imbalance. Hence, the results reported need to be interpreted considering these influential methodological aspects.

The main goal of this study was to design and validate a novel test battery and develop the most reliable statistical model for predicting working outcomes in trainee assistance dogs, i.e. estimating the probability of withdrawal either during training or during the first year of the working programme. This research builds upon the current state-of-the-art by combining subtests from previous research into a novel test

* Corresponding author.
E-mail address: marinara.marcato@tyndall.ie (M. Marcato).
[1] 0000-0002-7938-5227

https://doi.org/10.1016/j.applanim.2024.106212
Received 18 September 2023; Received in revised form 24 February 2024; Accepted 26 February 2024
Available online 28 February 2024
0168-1591/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

battery to assess a trainee dog's suitability for assistance work. The purpose of this study was to investigate the use of a novel test battery for the early prediction of future working outcomes in an assistance dog 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.

Previous research on test batteries for assessing the suitability of trainee dogs for assistance work has been extensively reviewed by Taylor and Mills (2006) and Marcato et al. (2022). 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. They will be thoroughly discussed in the methodology section in the context of the novel test battery. Taylor and Mills (2006) identified and discussed five key measures of quality of a test battery; namely, purpose, standardisation, reliability, validity, and practicality. Three factors were addressed in this research: purpose, standardisation and practicality.

Taylor and Mills (2006) and Brady et al. (2018) highlighted the importance of evaluating canine behaviour tests based on validity and reliability tests. This 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 (Batt et al., 2010. Graduating success (also referred to as training success or training outcome in literature) regards dogs who graduated at the end of training; it is the most commonly used measure because it requires less time to define. On the other hand, working success refers to 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 temporal together with circumstantial 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 a 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.

2. Methodology

2.1. Ethical approval

The novel ADTB subtest protocol developed 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). 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 implementation of the test battery.

2.2. Subjects

This study was part of a collaboration between Tyndall National Institute and the Irish Guide Dogs for the Blind (IGDB, Cork, Ireland). The subjects participating in this study are dogs who started the assistance dog training programme at the IGDB training centre between April/2019 and April/2021. The target number of subjects was 100 dogs (Laborde et al., 2017.

Dogs who were withdrawn from the programme due to health issues or selected as breeding stock were removed from the sample for further analysis. 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".

2.3. Test battery

The novel Assistance Dog Test Battery (ADTB) aims to differentiate dogs with different working outcomes. Hence, the subtests and ethogram were designed in line with previously published research results and adapted considering constraints, research goals and feedback from researchers and practitioners involved in 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 working outcomes. Two review papers present relevant research that supported the development of this novel test battery (Taylor and Mills, 2006; Marcato et al., 2022.

Taylor and Mills (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. Marcato et al. (2022) examined ten peer-reviewed studies on test batteries used to assess assistance dogs. Eight studies analysed the relationship between the items recorded and working outcomes. The items showing a statistically significant association with working outcomes were highlighted.

Table 1 provides details of the ADTB subtests' order, name, description, and references to the studies they were adapted from. This novel test battery comprised 14 subtests, which were presented to the subject dog in a controlled environment following a predetermined order. The subtests were combined into 5 parts with an overarching purpose: activity, responsiveness, sensitivity, focus, and emotional regulation. Part I "activity" captured the subject dog's response to a novel environment. Part II "responsiveness" was intended to gather inertial data for the posture recognition system and behavioural data for examining responsiveness to basic commands. Part III "sensitivity" evaluated body sensitivity and Part IV "focus" evaluated distractibility and ability to focus. Part V "emotional regulation" assessed traits such as fearfulness, aggressiveness, adaptability, and separation-related behaviours.

Key elements of 12 subtests presented in Taylor and Mills (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 and Mills (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., 2016. Elements from all the ten test batteries summarised in Marcato et al. (2022) 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 min. The subtests replicate a collection of typical real-life situations that 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 outcomes in the assistance dog programme existed.

An ethogram scoring system was developed to rate the dog's behavioural reaction considering previous research and dog trainers'

Table 1
Detailed description of the protocol for data collection using the ADTB.

Subtest (Duration)	Description (References)
Part I. Activity	
1. Familiarisation (1 min)	Handler and dog entered the test arena, handler let the dog off-leash, the dog was allowed to explore the arena for 1 min, handler called the dog back if they left their side (Wilsson and Sundgren, 1998
Part II. Responsiveness	
2. Walking (1 min)	Handler walked dog on a leash (Wilsson and Sinn, 2012
3. Standing (1 min)	Handler commanded the dog to "stand" and "wait"; if dog left the position, the handler would command them again until 1 min had elapsed (Harvey et al., 2016
4. Sitting (1 min)	Handler commanded the dog to "sit" and "wait"; if dog left the position, the handler would command them again until 1 min had elapsed (Harvey et al., 2016
5. Lying (1 min)	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 min had elapsed (Harvey et al., 2016
Part III. Sensitivity	
6. Body check (1 min)	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., 2016; Bray et al., 2019
7. Tea Towel (1 min)	Handler placed a quarter folded tea towel over the dog's back (off leash) while they were standing; handler waited for 5 s or until the dog removed the towel, repeating this two times (Harvey et al., 2016
Part IV. Focus	
8. Distractions (2 min)	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: (1) Teddy - teddy bear (2) Human - test leader greeted dog (Svartberg and Forkman, 2002; Bray et al., 2020 (3) Car - remote-controlled car moved along the path (4) Food - the bowl contained dog food (Harvey et al., 2016; Bray et al., 2017b
9. Kong (2 min)	(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 s (2) Interaction - The handler gives the Kong to the dog (off leash), and the dog is allowed to interact with Kong for 20 s (3) Return - The handler takes the Kong away from the dog (off leash) and returns it to storage (Batt et al., 2008; Tomkins et al., 2012; MacLean and Hare, 2018
10. Dog distraction (30 s)	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 (Batt et al., 2008; Tomkins et al., 2011; Bray et al., 2019
Part V. Emotional Regulation	
11. Crate (1 min)	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 min; handler opened the crate allowing the dog to come out (Svartberg, 2002; Batt et al., 2008; Tomkins et al., 2011
12. Petting (1 min)	Handler crouches down, freely petted and played with a dog (off leash) for about 20 s; 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 (1 min)	Handler left dog (off leash) alone in the test arena; both handler and test leader left room for 1 min; handler and test leader returned to room; handler greeted, petted and attached the leash on dog's collar (Wilsson and Sundgren, 1998; Bray et al., 2017b
14. Noise (1 min)	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 m away from and looking at them (Wilsson and Sundgren, 1997; Batt et al., 2008; Tomkins et al., 2011; Wilsson and Sinn, 2012; Asher et al., 2013; Bray et al., 2019
Estimated Total Duration: 16 min	

feedback. Table A.1 in the supplementary material 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.

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 research; however, the datasets were made publicly available on Mendeley Data (https://doi.org/10.17632/5mzfpt455r.1) (Marcato, 2023 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 and Mills (2006), including purpose and duration of the test, dog welfare concerns, and feasibility for practical use.

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 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. Fig. 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 the leash during subtests 1 and 13; hence, these are not indicated. The test arena was set up in a room mainly used for that purpose.

Figure A.1 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 criterion used for selecting the distraction dog related to their ability to settle quietly in a 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

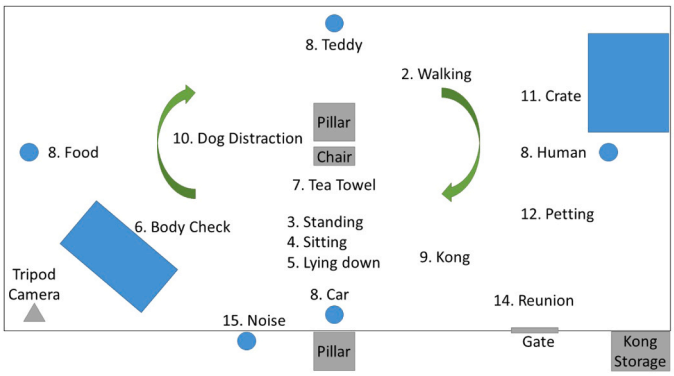


Fig. 1. Schematic representation of the test arena including key materials used and information about the test battery.

distraction dog was given a treat to chew on and was checked regularly while they stayed in the crate.

2.5. Data analysis

The data were analysed using R version 4.2.1 (Team, 2022. The code developed to implement the methodology which yielded the results presented in this study was also made publicly available on GitHub (link). GLMs 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, service 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 A.1).

Composite scores were calculated based on ethogram items in order to combine information from multiple items into one variable. Table A.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 the methodology presented in Sections 2.5.1 and 2.5.2.

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 working 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 $p < 0.15$.

2.5.2. Predictability

The resulting datasets consisted of 111 ADTB ethogram variables, including 70 original items (53 ordinal and 17 categorical) and 41 composite scores 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. (2016). 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 ($p < 0.15$), the variable with the lowest p-value was retained. This approach was adopted to select covariates more strongly related to working 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 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". A backward feature selection algorithm was implemented using Recursive Feature Elimination (RFE) to maximise the f1-score, which is defined as the harmonic average of the precision and recall. 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. Model fitness was analysed using the likelihood ratio test for nested models, which follows a chi-squared distribution. The area under the ROC curve was selected to compare the performance of different models, as it is a more appropriate measure for imbalanced datasets (Luque et al., 2019; Haixiang et al., 2017.

Conclusively, while the f1-score was used to optimise the model, 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.

3. Results

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 (Mean = 15.3, SD = 2.04). Their age at DC1 varied from 13 to 23 months (Mean = 16.15, SD = 2.01).

The failure rate in this sample was 21%, as 21 dogs failed and 77 dogs succeeded in their training. 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.

The descriptive statistics of ordinal ethogram variables are shown in Table B.1 and categorical variables in Tables B.2 and B.3. The marking scale used to score the items is described in Table A.1.

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 B.4, 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 B.5 and B.6.

Two ethogram variables showed a significant association with working outcomes at traditional levels ($p < 0.05$): the ethogram item "Petting Confidence During" ($p = 0.03$) and its derived composite score "S Petting Stimulus During_mean" ($p = 0.02$). 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" to reflect the engagement between handler and dog, as shown in Table A.2.

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 B.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 1 and the lines delineate groups of variables that are related to each other, as indicated in Table A.2.

Step 1 removed features that had $p > 0.15$ in the univariate logistic regression analysis as shown in Tables B.4, B.5, and B.6. It selected 15 out of the 111 ADTB 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. No variables were removed as none exceeded the threshold, as shown in the correlation matrix in Figure B.1.

Hence, seven features were inputted to the backward feature selection algorithm, which was set to maximise f1-score using LOOCV to estimate performance metrics. Table 2 shows the model parameters for the four covariates selected for the optimal multivariate model; namely, "S Sensitivity_mean", "S Distractions Pull_mean", "S Crate Stimulus_mean", and "S Petting Stimulus During_mean".

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".

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 A.2, then the ethogram item scores in Table A.1. Table B.8 combines key information from both tables for ease of reference. Figures B.2 to B.8 show the variables selected for the reduced dataset created with DC1 data.

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 1 "walking"; 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".

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 $P(\tilde{\chi}_4^2 \geq 20.00) = 5.00e^{-04}$.

Table 3 shows the performance metrics estimated from the predictions made by the optimal model, with four composite scores derived from the ethogram items using LOOCV. Table B.9 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 2a show the probability of failure predicted by the best logistic regression model for all the 98 dogs in the dataset. In particular, Fig. 2a displays the working outcome obtained by each dog, i.e.,

Table 2
Optimal multivariate logistic regression models selected by RFE.

Variable	Estimate	P-value	OR	2.5%	97.5%
Data Collection 1					
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
Data Collection 2					
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 3
Performance metrics from the models created with data from DC1 and 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

whether the dog failed or succeed at graduating and working for one year. The lowest probability of failure belonged to guide dogs (Mean = 0.14, SD = 0.13), service dogs (Mean = 0.24, SD = 0.24), and withdrawn dogs (Mean = 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).

3.2. Data Collection 2

The sample comprised a total of 60 dogs, including 32 females (53%) and 28 males (47%); there were 25 LR, 34 LRxGR, and 1 GR. Their age at arrival ranged from 13 to 22 months (Mean = 15.3, SD = 1.9). Their age at DC2 varied from 15 to 24 months (Mean = 17.8, SD = 1.8). The sample size was smaller in DC2 compared to DC1 because of the global pandemic, as the data collection had to be temporarily cancelled.

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

The descriptive statistics of ordinal ethogram variables are shown in Table C.1 and categorical variables in Tables C.2 and C.3. The marking scale used to score the items is described in Table A.1.

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.4, 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.5 and C.6.

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".

3.2.2. Predictability

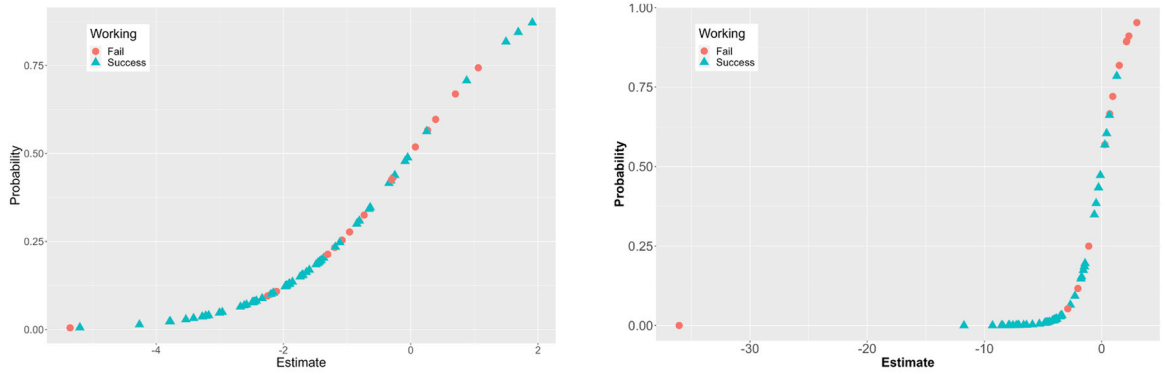
A three-step feature selection strategy was carried out. The initial dataset contained two demographic variables; 70 original items and 41 composite scores. Table C.7 indicates which of the variables selected in step 1 satisfied steps 2 and 3, and were selected by the RFE algorithm. They were arranged according to the order of the subtests as shown in Table 1 and the lines separate groups of variables that are related to each other as indicated in Table A.2.

Step 1 removed features that had $p > 0.15$ in the univariate logistic regression analysis as shown in Tables C.4, C.5 and C.6. 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 C.1 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 inputted to the backward feature selection algorithm which was set to maximise f1-score using LOOCV to estimate performance metrics. Table 2 shows the model parameters for the six covariates selected for the optimal multivariate model; namely, "Distractions First Response Car", "S Walking Pull_mean", "S Familiarisation Handler_mean", "S Isolation Stimulus", "S Crate Handler_mean", and "S Sensitivity".

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 11 "crate".

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 A.2, then the ethogram item scores in Table A.1. Table C.8 combines key information from both tables for ease of reference. Figures C.2 to C.10 show the variables selected for the



(a) DC1 dataset. (b) DC2 dataset.

Fig. 2. Probability of failure predicted by the best logistic regression models per true working outcome (fail, success).

reduced dataset created with DC2 data.

The probability of failure decreased as the value of the following scores increased: "S Familiarisation Handler_mean", "S Walking Pull_mean", "S Sensitivity_mean", "S Crate Handler_mean", "S Isolation Stimulus". 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"; were frequently actively seeking attention, whining or barking in the subtest 11 "crate"; and more frequently explored and paces around the test arena during the subtest 13 "isolation".

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 $P(\tilde{\chi}_6^2 \geq 30.99) = 2.54e^{-05}$.

Table 3 shows the performance metrics estimated from the predictions made by the optimal model with one ethogram item and five composite scores using LOOCV. Table C.9 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 Familiarisation Handler_mean", "S Isolation Stimulus", "S Crate Handler_mean", and "S Sensitivity_mean".

Figure 2b show the probability of failure predicted by the best logistic regression model for all the 60 dogs in the dataset. In particular, Fig. 2b displays the working outcome obtained by each dog, i.e., whether the dog failed or succeed at graduating and working for one year. The lowest probability of failure belonged to guide dogs (Mean = 0.04, SD = 0.10), service dogs (Mean = 0.22, SD = 0.25), and withdrawn dogs (Mean = 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 service dogs (p = 6.9e-04), and service and withdrawn dogs (p = 2.2e-03) were highly significant.

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 indicated to be a more definitive measure of true success than training outcomes (Batt et al., 2010. Previous research has 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 In For Training (IFT) (Bray et al., 2019, 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 area under the 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.

Section 4.1 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 4.2 discusses the results achieved in this study and previously published research.

4.1. 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 shows the 15 variables selected in DC1 and 23 in DC2.

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

Table 4
Ethogram items and composite scores are separated per subtest and by Data Collection (DC1, 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** * S Familiarisation Handler_prod *
2. Walking		**S Walking Pull_mean***** S Walking Pull_prod *** S Walking Distractions Pull_mean ** Walking Pull on leash *** Walking Pull strength ***
3. Standing		
4. Sitting		
5. Lying down		
6. Body check	Table ** **S Sensitivity_mean** **	Table * **S Sensitivity_mean** **
7. Tea towel	First Response Indifferent *	Second Response Indifferent ** Second Response Change from Neutral *
8. Distractions	First Response Human * **S Distractions Pull_mean** ** S Distractions Pull_prod **	**First Response Car** ***
9. Kong		Interaction Response to stimulus *** Return Handler ** S Kong Response_prod *
10. Dog distraction		
11. Crate	Actively Seeking Attention * Sniffing Exploring ** **S Crate Stimulus_mean****	Actively Seeking Attention ** Whining * **S Crate Handler_mean**** S Crate Handler_prod *
12. Petting	Handler Stimulus * Confidence During *** **S Petting Stimulus During_mean** ** S Petting Stimulus During_prod ** S Sociability_mean ** S Sociability_prod *	
13. Isolation		Response Time oriented *** Response Exploration *** **S Isolation Stimulus** *** S Isolation Handler_mean ***
14. Noise		

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

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 3 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). 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.

4.2. Comparison with previous research

Two types of comparison are discussed. In Section 4.2.1, 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 4.2.2, 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 organised in a table and discussed.

4.2.1. 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 and Sinn (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. (2016) 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 3–5 were similar to the "obedience training" in Weiss and Greenberg (1997) and Weiss (2002), to the "novel situation" test in Weiss (2002), and to the "obedience" subtest in Harvey et al. (2016). 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 Marcato et al. (2023). Weiss and Greenberg (1997) and Weiss (2002) did not evaluate the link between behavioural measures related to obedience and training outcomes. Harvey et al. (2016) 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. (2016) 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. (2016). 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. (2016). 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 and Hare (2018) Batt et al. (2008), Tomkins et al. (2012), and MacLean and Hare (2018). Dogs who used both paws more frequently (Batt et al., 2008 or preferred using the right paw Tomkins et al. (2012) were more successful. MacLean and Hare (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 Batt et al. (2008), 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. This result is in line with Tomkins et al. (2011) and contrary to Batt et al. (2008) 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. 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.

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 and Sundgren (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" have been performed in many studies, only Batt et al. (2008), 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 Batt et al. (2008) and Asher et al. (2013) reported no statistical significance, while 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 (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.

4.2.2. Prediction performance

Table 5 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.

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. Batt et al. (2008) 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. (2016) 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

Table 5
Comparison between the performance results achieved by the two classifiers in this study and previously reported research.

Test Name / Publication	N	Imb.	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 Batt et al. (2008)	43	0.53	6						0.27
	43	0.53	14						1.00
	43	0.53	14–20						0.32
PPA Asher et al. (2013)	465	0.64	2			0.93	0.08	1.00	
JGDBT Harvey et al. (2016)	73	0.71	5						0.80
	72	0.79	8						0.87
YAT Bray et al. (2017a)	98	0.68	14–17		0.75				
DCTB MacLean and Hare (2018)	111	0.68	24	4-fold	0.76				0.72
	121	0.77	24	Hold out	0.68				0.61
IFT Bray et al. (2019)	5967	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. Imb = Imbalance: the proportion of examples from the majority class (success). Age: given in months. Method: Leave One Out Cross-Validation.

and Hare (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 the 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.

5. Conclusion

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 study was to design and validate a new test battery to estimate the suitability of trainee dogs for assistance work at different stages of the training programme. Hence, the novel 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 an area under the ROC of 0.74, the DC2 model outperformed previous studies on training outcome prediction using test batteries with an area under the 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. Data was made available through a public repository following the FAIR principles (Wilkinson et al., 2016. 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 research 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.

Funding

This publication has emanated from research supported in part by a grant from the European Regional Development Fund (ERDF) through the Ireland-Wales INTERREG Programme under the Celtic Advanced Life Science Innovation Network (CALIN) project [grant number]; from Science Foundation Ireland (SFI) [INSIGHT grant number] which is co-funded under the European Regional Development Fund; from SFI and the Department of Agriculture, Food and Marine [VistaMilk grant number]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

CRediT authorship contribution statement

Marinara Marcato: Writing – review & editing, Writing – original draft, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Salvatore Tedesco:** Writing – review & editing, Supervision, Formal analysis. **Conor O′ Mahony:** Writing – review & editing, Supervision. **Brendan O′Flynn:** Writing – review & editing, Supervision. **Paul Galvin:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This research is related to the PhD of Marinara Marcato, which is part of collaborations with the Irish Guide Dogs for the Blind.

Acknowledgements

We would like to thank the undergraduate students Maria Vittoria Lorenzi, Hazel Craven, Conghal Hewson, and Jennifer Kenny, and the staff at the IGDB for their contribution to the data collection.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.applanim.2024.106212.

References

ADI, 2022.ADI Terms & Definitions.⟨https://assistancedogsinternational.org/resources/adi-terms-definitions/⟩.

Asher, L., Blythe, S., Roberts, R., Toothill, L., Craigon, P.J., Evans, K.M., Green, M.J., England, G.C., 2013. A standardized behavior test for potential guide dog puppies: Methods and association with subsequent success in guide dog training. J. Vet. Behav. 8, 431–438. https://doi.org/10.1016/j.jveb.2013.08.004.

Batt, L., Batt, M., Baguley, J., McGreevy, P., 2010. Relationships between puppy management practices and reported measures of success in guide dog training. J. Vet. Behav.: Clin. Appl. Res. 5, 240–246. https://doi.org/10.1016/j.jveb.2010.02.004.

Batt, L., Batt, M.S., Baguley, J.A., McGreevy, P.D., 2008. Factors associated with success in guide dog training. J. Vet. Behav.: Clin. Appl. Res. 3, 143–151. https://doi.org/10.1016/j.jveb.2008.04.003.

Brady, K., Cracknell, N., Zulch, H., Mills, D.S., 2018. A systematic review of the reliability and validity of behavioural tests used to assess behavioural characteristics important in working dogs. Front. Vet. Sci. 5 https://doi.org/10.3389/fvets.2018.00103.

Bray, E.E., Gruen, M.E., Gnanadesikan, G.E., Horschler, D.J., Levy, K.M., Kennedy, B.S., Hare, B.A., MacLean, E.L., 2020. Cognitive characteristics of 8- to 10-week-old assistance dog puppies. Anim. Behav. 166, 193–206. https://doi.org/10.1016/j.anbehav.2020.05.019.

Bray, E.E., Gruen, M.E., Gnanadesikan, G.E., Horschler, D.J., Levy, K.M., Kennedy, B.S., Hare, B.A., MacLean, E.L., 2021. Dog cognitive development: A longitudinal study across the first 2 years of life. Anim. Cogn. 24, 311–328. https://doi.org/10.1007/s10071-020-01443-7.

Bray, E.E., Levy, K.M., Kennedy, B.S., Duffy, D.L., Serpell, J.A., MacLean, E.L., 2019. Predictive models of assistance dog training outcomes using the canine behavioral assessment and research questionnaire and a standardized temperament evaluation. Front. Vet. Sci. 6, 49. https://doi.org/10.3389/fvets.2019.00049.

Bray, E.E., Sammel, M.D., Cheney, D.L., Serpell, J.A., Seyfarth, R.M., 2017a. Effects of maternal investment, temperament, and cognition on guide dog success. Proc. Natl. Acad. Sci. USA 114, 9128–9133. https://doi.org/10.1073/pnas.1704303114.

Bray, E.E., Sammel, M.D., Seyfarth, R.M., Serpell, J.A., Cheney, D.L., 2017b. Temperament and problem solving in a population of adolescent guide dogs. Anim. Cogn. 20, 923–939. https://doi.org/10.1007/s10071-017-1112-8.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: review of methods and applications. Expert Syst. Appl. 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035.

Harvey, N.D., Craigon, P.J., Sommerville, R., McMillan, C., Green, M., England, G.C.W., Asher, L., 2016. Test-retest reliability and predictive validity of a juvenile guide dog behavior test. J. Vet. Behav.: Clin. Appl. Res. 11, 65–76. https://doi.org/10.1016/j.jveb.2015.09.005.

Hosmer, D., Lemeshow, S., Sturdivant, R., 2013. Applied logistic regression. Wiley Series in Probability and Statistics. Wiley.

IGDB, 2021.Irish Guide Dogs for the Blind | Guide and Assistance Dogs.⟨https://www.guidedogs.ie/⟩.
Jones, A.C., Gosling, S.D., 2005. Temperament and personality in dogs (Canis familiaris): A review and evaluation of past research. Appl. Anim. Behav. Sci. 95, 1–53. https://doi.org/10.1016/j.applanim.2005.04.008.
Laborde, S., Mosley, E., Thayer, J.F., 2017. Heart Rate variability and cardiac vagal tone in psychophysiological research – recommendations for experiment planning, data analysis, and data reporting. Front. Psychol. 08 https://doi.org/10.3389/fpsyg.2017.00213.
Luque, A., Carrasco, A., Martín, A., de las Heras, A., 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognit. 91, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023.
MacLean, E.L., Hare, B., 2018. Enhanced Selection of Assistance and Explosive Detection Dogs Using Cognitive Measures. Front. Vet. Sci. 5, 236. https://doi.org/10.3389/fvets.2018.00236.
Marcato, M., 2023.Inertial sensor dataset for Dog Posture Recognition.10.17632/MPPH6BMN7G.1.
Marcato, M., Kenny, J., O'Riordan, R., O'Mahony, C., O'Flynn, B., Galvin, P., 2022. Assistance dog selection and performance assessment methods using behavioural and physiological tools and devices. Appl. Anim. Behav. Sci. 254, 105691 https://doi.org/10.1016/j.applanim.2022.105691.
Marcato, M., Tedesco, S., O'Mahony, C., O'Flynn, B., Galvin, P., 2023. Machine learning based canine posture estimation using inertial data. PLOS ONE 18, e0286311. https://doi.org/10.1371/journal.pone.0286311.
Svartberg, K., 2002. Shyness–boldness predicts performance in working dogs. Appl. Anim. Behav. Sci. 79, 157–174. https://doi.org/10.1016/S0168-1591(02)00120-X.
Svartberg, K., Forkman, B., 2002. Personality traits in the domestic dog (Canis familiaris). Appl. Anim. Behav. Sci. 79, 133–155. https://doi.org/10.1016/S0168-1591(02)00121-1.
Taylor, K.D., Mills, D.S., 2006. The development and assessment of temperament tests for adult companion dogs. J. Vet. Behav. 1, 94–108. https://doi.org/10.1016/j.jveb.2006.09.002.
Team, R.C., 2022.R: A Language and Environment for Statistical Computing.R Foundation for Statistical Computing.
Tomkins, L.M., Thomson, P.C., McGreevy, P.D., 2011. Behavioral and physiological predictors of guide dog success. J. Vet. Behav.: Clin. Appl. Res. 6, 178–187. https://doi.org/10.1016/j.jveb.2010.12.002.
Tomkins, L.M., Thomson, P.C., McGreevy, P.D., 2012. Associations between motor, sensory and structural lateralisation and guide dog success. Vet. J. 192, 359–367. https://doi.org/10.1016/j.tvjl.2011.09.010.
Weiss, E., 2002. Selecting Shelter Dogs for Service Dog Training. J. Appl. Anim. Welf. Sci. 5, 43–62. https://doi.org/10.1207/S15327604JAWS0501_4.
Weiss, E., Greenberg, G., 1997. Service dog selection tests: Effectiveness for dogs from animal shelters. Appl. Anim. Behav. Sci. 53, 297–308. https://doi.org/10.1016/S0168-1591(96)01176-8.
Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, 3, 10.1038/sdata.2016.18..
Wilsson, E., Sinn, D.L., 2012. Are there differences between behavioral measurement methods? A comparison of the predictive validity of two ratings methods in a working dog program. Appl. Anim. Behav. Sci. 141, 158–172. https://doi.org/10.1016/j.applanim.2012.08.012.
Wilsson, E., Sundgren, P.E., 1997. 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. Appl. Anim. Behav. Sci. 53, 279–295. https://doi.org/10.1016/s0168-1591(96)01174-4.
Wilsson, E., Sundgren, P.E., 1998. Behaviour test for eight-week old puppies—heritabilities of tested behaviour traits and its correspondence to later behaviour. Appl. Anim. Behav. Sci. 58, 151–162. https://doi.org/10.1016/S0168-1591(97)00093-2.



Marinara Marcato is a PhD candidate at Tyndall National Institute. She has been working on a collaborative project entitled Guide Dog 4.0 with Swansea University and the Irish Guide Dogs for the Blind (IGDB). Her goal is to develop performance assessment techniques for guide and assistance dogs through the deployment of technology. Her research interests lie in the areas of artificial intelligence and wearable devices. Prior to joining Tyndall, Marinara earned her B.E. Hons Electrical Engineering degree from Londrina State University.



Dr. Salvatore Tedesco (Member, IEEE) is a Senior Researcher and a Team Leader at Tyndall National Institute, Cork, Ireland. He is responsible for leading a research team in the wearable and data analytics area focused on his main research interests on wearable technologies for healthcare and well-being, human motion analysis in sports and clinical populations, digital health, physiological monitoring, signal processing, edge analytics, and machine learning.



Dr. Conor O'Mahony (M. Eng. Sc '01, PhD '04) is a Principal Scientist with the Tyndall National Institute, Cork, Ireland. His team is focused on the development of Micro Transdermal Interface Platforms (MicroTIPs) - high-value, wearable systems that with applications in transdermal drug delivery, diagnostics and intelligent woundcare. He has published over sixty peer-reviewed journal articles and is a Senior Member of the IEEE.



Dr. Brendan O'Flynn is Head of Group for Wireless Sensor Networks with the Tyndall National Institute, Cork, Ireland. His group is focused on the development of smart, autonomous, and reconfigurable wireless sensing systems and the creation of information from data.



Dr. Paul Galvin is Head of the ICT for Health Strategic Programmes, Head of the Life Sciences Interface Group and Head of the Bioelectronics Cluster at the Tyndall National Institute, where he drives an "ABC ecosystem" through collaborations with leading academic researchers, global leaders in the medtech and pharmaceutical industries, and with clinical experts. He has published research articles in leading international peer-reviewed journals across a range of scientific disciplines.