What Makes a Good Service Dog?
An Exploration of Temperament and Learning Models in Predicting Training Success

By: Aarushi Adlakha, Anne Townsend, Kalena Bing, Katie Bausenwein

Introduction

Coach, the Princeton Public Safety service dog, is often spotted at orientation, sporting events, on her walks, and at many other campus activities. Over the years, members of our group have had several interactions with her and we've all immediately noticed her friendly personality and excitement to meet students. More infrequently, the residential colleges have Newfoundland therapy dogs visit the students as well. These dogs tend to be mellow and do not get rewarded with food until the end of their two hour visits, two features that are very distinct from Coach. These differences made our group wonder what truly makes a good service dog? Could it be differences in breed, temperament, environmental exposures, training routines, or some combination of all of these?

To note, although terms like therapy dog, service dog, and emotional support animal are often used interchangeably, our group adopts the ADA definition of a service dog, as it provides the most appropriate framework for our modeling and research aims[1].

We found the question of 'what makes a good service dog' particularly important to those training and utilizing service dogs, as it costs $45,000-70,000 to train one (Cleghern et al., 2018). Additionally, about 50% of service dogs fail out of training–with the majority of reasons being behavioral issues (Duffy and Serpell, 2012, Harvey et al., 2017). These dogs are a huge investment, yet half of them do not even make it to the end of their training. Some of these dogs could have failed out of their program at later points, meaning a significant amount of money was already spent on training that did not work out. For these reasons, it would be ideal to have a tool which could assess the likelihood that the dog could successfully complete training.

This becomes a very difficult task when considering that training methods and final requirements for service dogs are not standardized (Acebes et al., 2022). Without this standardization, it becomes unclear what traits should be emphasized in these predictive assessments. How can a model be built without a clear end goal? Likely because of the variation in training, there has also been very limited or hyperspecific research on service dogs. Papers are often focused on subcategories, such as police dogs, emotional support animals, detection dogs, etc.–all of which have different behavioral goals and preferred temperaments. Within these categories, there have been some signs of success in creating assessment tools with the use of machine learning and other computational methods. However, the scope of these tools are restricted to specific aforementioned subcategories, and not applicable across occupation.

In this paper, we argue that breed-related temperament profiles–shaped by both neural architecture and behavioral specializations–play a critical role in service dog success, and we demonstrate how incorporating these traits into computational learning models can help address key gaps in the literature on predicting and optimizing training outcomes.

Temperament

A growing body of research has emphasized that temperament is not a superficial or cosmetic aspect of canine behavior, but rather a deep, neurodevelopmentally grounded set of

[1] Service animals are defined as dogs that are individually trained to do work or perform tasks for people with disabilities (ADA, 2020). Further clarification on differences in training and responsibility of the three types of working animals can be found at: https://www.mncanineconsulting.com/post/service-dogs-esd-therapy-dog

traits that vary systematically across breeds. In particular, Hecht et al. (2021) provides one of the most comprehensive and biologically informed explanations for this variability. The study, which integrates data from the C-BARQ[2], questionnaire with MRI-based neuroanatomical analyses across dozens of dog breeds, argues that differences associated with body size is a major driver of temperament profiles across breeds.

A key conceptual contribution of Hecht et al. is that they frame these temperament variations as part of a broader pattern of evolution seen across mammals: as body size increases, developmental timelines lengthen, and neural systems mature differently. One of the core discoveries of the MRI analyses is that brain regions do not scale uniformly as overall brain size changes. Instead, certain structures expand or contract disproportionately relative to total brain volume. While smaller-bodied breeds tend to have relatively larger limbic regions (amygdala, hippocampus) when corrected for whole-brain size, larger-bodied breeds show proportionately larger cortical association regions, especially in frontal and parietal areas. This has enormous implications for temperament. Larger amygdala volume is associated with higher fear responses, increased vigilance, and defensive aggression. This aligns with the C-BARQ data, where smaller breeds scored significantly higher on fear, excitability, and stranger-directed aggression. This suggests a structural and neurobiological explanation for why many small breeds show "reactive" temperaments–the fear circuitry occupies a larger share of the brain's total resources, increasing sensitivity to threat and novelty. Conversely, larger breeds (with a relatively smaller proportional amygdala) often display more behavioral stability and reduced general fear, which are critical traits for service work. Larger-bodied breeds also have relatively expanded cortical areas which are involved in executive control, social cognition, sensory integration, and emotional modulation. This allows for attributes like better impulse control, improved adaptability to novelty, and reduced reactivity.

These results help explain why training alone cannot fully eliminate certain tendencies, as they are not quirks but neurodevelopmentally embedded features of the breed. Temperament is not merely a product of training or environment, but emerges much earlier from stable, inherited differences. Together, this justifies using temperament assessments (like C-BARQ) and breed-informed selection practices when identifying service dog candidates. The qualities that make a dog safe, stable, and effective in service contexts, such as low anxiety, flexible emotional regulation, and reliable social engagement, are linked to identifiable brain structures that vary by breed and size. Selecting good service dogs requires an understanding of these underlying neural differences, and helps explain why some dogs naturally thrive in therapeutic work while others struggle despite training (like Labradors, Retrievers, or Poodles). The goal is not to rely on breed stereotypes but to acknowledge these neurodevelopmental realities that shape behavioral tendencies and training success.

It is also important to consider a major limitation to the Hecht et al. study, in that it relies on breed-level averages rather than individual-level neurobehavioral data. Even though certain breeds or size classes show characteristic patterns, any given dog carries its own unique developmental history shaped by early socialization, trauma, prior training, health status, and environmental enrichment. These experiential factors can strongly influence temperament, often

[2] The C-BARQ questionnaire consists of 100 questions in 7 distinct categories: 8 questions related to training, 27 to aggression, 18 to fear, 8 to separation, 6 to excitability, 6 to attachment and attention seeking, and 27 to miscellaneous. Notably, the C-BARQ primarily measures problematic behavioral traits. These are coded in a Likert scale from 0 to 4 where higher C-BARQ scores are indicative of the behavior being less desirable (Zapata et al., 2022).

in ways that override or interact with breed predispositions. For example, a Golden Retriever with a history of chronic stress or poor socialization may be far less suitable for service work than a mixed-breed shelter dog with a stable upbringing and strong social tendencies. The Hecht et al. findings should therefore be interpreted as identifying probabilistic tendencies, not deterministic predictors of behavior.

To reiterate, this study compares breeds rather than individuals, and is not designed to answer the question most relevant to service dog programs: Which specific dogs will succeed in training? Functional MRI, in theory, could offer more direct measures of neural responsiveness to social or emotional stimuli, but in practice, its application is limited. Awake fMRI studies in dogs are logistically demanding, limited to highly trainable individuals, and typically exclude the very populations (such as shelter dogs) whose success prediction would be most valuable. In this sense, neuroimaging contributes useful theory by showing that temperament differences have biological underpinnings, but it is not a practical method for predicting service dog success on a dog-by-dog basis, especially in diverse populations that include mixed breeds and shelter dogs.

However, we found a statistical method that provides accurate predictions across intrabreed differences, called Latent Class Analysis (LCA). In this case, LCA will be applied to C-BARQ results. LCA groups data based on similar trends in answer selection, forming the "classes." Every class has a proportional profile for the likelihood of selecting each answer in each C-BARQ question. Each subject will be grouped into one of these classes, based on which of the probability profiles their answers align with the closest. What qualities are shared amongst the data (forming classes), must be interpreted. This can be done through comparing the proportional profiles–assessing which questions had the greatest difference in a typical score. Through looking at the topic of these questions and how they differed, the class can be assigned a title which carries meaningful information about how it is different from the others (Weller et al., 2020; Zapata et al., 2022).

Zapata et al. used LCA to identify dog temperament classes from C-BARQ data. They (arbitrarily) preset the number of classes to three and ran this analysis on the C-BARQ results from 57,454 dogs of 350 different breeds (2022). Through comparing the questions with the largest difference in score between classes, the resulting temperament groups were identified as fearful, calm, and aggressive (Zapata et al., 2022). Within the resulting classes, they were able to identify intrabreed differences. For example, out of the population of Shih Zus, 27% were calm, 41.6% were fearful, and 31.4% were aggressive (Zapata et al., 2022). This shows that within one breed, individual dogs are not guaranteed to have the same temperament, even with the previously discussed anatomical predisposition. Thus, it is important to assess each individual dog to truly understand their temperament and chance at success in becoming a service dog.

The methodology of using LCA on the C-BARQ data could be slightly tailored to achieve this predictive goal. If the ideal number of classes was selected, then all of these aforementioned profiles could be used as a reference to sort any individual dog's C-BARQ score into a temperament category. Remembering that behavioral issues are the main reason dogs do not make it through their training, this would be very valuable information to tailor methods or discontinue with the dog before too much money is spent.

Next, we consider how these findings apply to service-dog training and how temperament-based differences can be incorporated into models that predict training success.

Learning Models in Training Service Dogs

Training a service dog involves a sequence of increasingly complex behaviors that must be acquired, chained together, and generalized across diverse contexts. Professional trainers typically approach training a new dog by breaking down the overall task into smaller stages, such as obedience, detection, socialization, and task-specific behaviors, and teach each stage separately, gradually integrating them into a coherent behavioral repertoire (*From puppy to partner: How service dogs are trained* 2025). This multi-stage process closely aligns with the learning frameworks discussed in class, in using successive approximations to shape desired responses.

For behaviors such as detection of physiological changes, like detecting gluten or fluctuations in blood glucose, service dogs must first learn to associate a particular cue, internal or external, such as odor signature, chemical shift, or behavioral signal, with an outcome. This mirrors classical conditioning, in which the cue acquires value through repeated pairings with a biologically relevant consequence (reward, trainer feedback, or the target stimulus itself). These cues then become the conditioned stimuli that reliably elicit attentional or searching behaviors. In contrast, task training, often including item retrieval, opening or closing doors, or navigating to specific locations, requires the dog to learn action-outcome pairings, which are best captured by operant conditioning and reinforcement learning (RL) models. Each action (e.g., pulling a rope, closing a fridge door, returning to the handler) is shaped through reinforcement, with positive outcomes increasing the likelihood of performing that action in the future. Moreover, because many of these tasks involve long action sequences with intermediate decisions, models like SARSA help formally describe how dogs incrementally update action values and prediction errors as they experience the consequences of each step in real time, as we elaborate further in the next section.

A challenge in developing a unified theoretical account of service-dog learning is the limited empirical literature directly applying formal learning models to service-dog training. Existing studies often focus on isolated components, like preference for praise vs. food vs. gestures (Fukuzawa & Hayashi, 2013) rather than multi-stage learning. Given this gap, the framework we present relies on extrapolating concepts from class materials and well-established animal learning theories into the domain of service-dog training.

Rescorla - Wagner

The RW model can help explain how a service dog learns which cues matter by formalizing how the dog assigns predictive value to a cue through repeated experience. It is important to note that RW models how the significance and meaning of a cue updates over time, rather than modeling the dog's behavioral response to the cue. For example, a diabetic-alert-dog (DAD) is repeatedly presented with saliva samples in training that were collected from their human during episodes of low blood sugar. Then, the dog receives a reward immediately after sniffing these low-glucose samples. These trials set the outcome value (λ) to 1, whereas trials with normal glucose samples set λ to 0. With each presentation, the dog compares the actual outcome (reward vs. no reward) to its current expectation, updating the associative strength of the odor cue according to the RW equation:

$$V_{new} = V_{old} + \eta(\lambda - V_{total})$$

Because the odor is the only cue in this example, V_{total} equals the odor's own V value. Over many iterations, the prediction error shrinks and a DAD's internal representation of the odor becomes strongly predictive of low blood sugar, essentially learning that this smell means I

should pay attention. However, in other service work, dogs may encounter multiple cues present at the same time, such as odor, changes in the handler's posture, or subtle shifts in breathing. In this case, their associative strengths add together to form V_{total}. This means the dog is not evaluating any cue in isolation but summing the predictive value of all available signals to generate its expectation of an outcome. Breed or temperament related differences in learning rate (η) could influence how quickly or strongly dogs acquire this association, connecting directly to the temperament findings examined earlier. Once the odor acquires predictive value through RW, operant conditioning then shapes the dog's behavioral response, teaching what to do when this now-meaningful cue is detected.

SARSA

SARSA (State-Action-Reward-State-Action) is an on-policy, temporal difference (TD) reinforcement learning algorithm, in which the action-value function is updated based on the action actually taken by the learner under its current policy behavior. The model estimates the value of the state-action pairs Q (s,a) and updates these estimates after every transition experienced by the agent *(PSY338, Week 5-6)*:

$$PE(S_{t+1}) = R(S_{t+1}) + Q(S_{t+1}, a_{t+1}) - Q(S_t, a_t)$$
$$Q_{new}(S_t , a_t) = Q_{old}(S_t , a_t) + \eta \cdot PE$$

Task training a service dog is a multi state process, involving multiple sequential decision making tasks. For example, in a task requiring it to fetch insulin from the fridge for a diabetic patient, the following subtasks form the complete task:

- Go to the fridge from wherever you are
- Pick up a rope connected to the fridge
- Walk backwards with the rope
- Find and grab item in the fridge
- Walk over to owner and place in hand
- Go back to the fridge
- Close the door with nose or paw

At each step, the dog is at a distinct state S_t, selects an action a_t, and receives a reinforcement $R(S_{t+1})$. The trainers' verbal praise, food reward etc. constitutes the reward signal. The subsequent shape and action form the basis for updating Q (S, A).

Applying SARSA to the aforementioned task, the dog must learn how to navigate to the refrigerator as the first step in a multi-action behavioral chain. During early training trials, the dog takes the *red path* to the fridge (Fig. 1). Upon successfully reaching the fridge, it receives a reward R, producing a positive temporal-difference (TD) prediction error. Under SARSA, this reward updates the action-value estimate for the state–action pair (S_t,A_t) corresponding to "walk to fridge (red path)" according to the update rule. Because the reward at the fridge leads to a successor action with positive value, $Q(S_{t+1},A_{t+1})$, the Q-value for the red-path action increases, making it more likely for the dog to repeat this route.

After several successful trials leading to a steady increase in the Q-value for the red path, an obstacle is introduced along the red path (Fig. 2). When the dog attempts the previously learned action, it reaches a blocked state and fails to receive the expected reward, generating a negative TD error. SARSA therefore reduces the Q-value for "walk straight (red path)" because the actual next action and its resulting state did not result in a reward. The dog then explores an

alternative route, which is depicted by the *green path*, and upon successfully reaching the fridge and obtaining a reward, the corresponding Q-value for that new state–action pair increases.

This illustrates SARSA's core property: the value updates depend on the actions the dog actually takes (including inefficient or exploratory ones) rather than the hypothetical optimal action. As a result, the dog adaptively shifts its behavior toward the path that truly produces reward in the current environment.

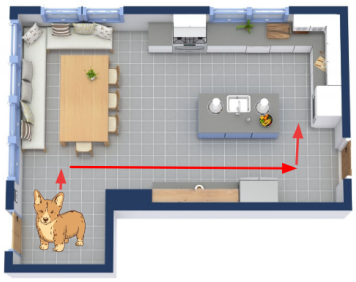

Fig 1: The dog under training learns that the red path is an optimal route to the fridge.

Fig 2: Illustration of SARSA updating when the dog encounters a barrier.

Together, both classical and instrumental learning models help illustrate how service dog training is a complex, layered process; and that learning can be captured by a variety of theoretical frameworks that capture different components of behavior acquisition. Classical conditioning (for instance, the Rescorla Wagner model) explains how dogs learn which cues matter by assigning predictive values to odors, physiological signals, or social cues. Operant conditioning and temporal difference reinforcement learning models (like SARSA), capture how dogs learn what to do once those cues have meaning. Importantly, the impact of breed and temperament differences can be understood as being nested within these computational frameworks through their influence on the dog's learning rate (η) parameter. The learning rate (η) determines how rapidly new associations are formed in R-W, and how new experiences impact action-value estimates in SARSA. Temperament linked traits such as impulsivity, attentional control, persistence, or anxiety could systematically alter the effective learning rate. For instance, impulsive or highly exploratory dogs may weigh recent prediction errors more heavily, producing a higher η and therefore faster, but potentially noisier, learning. Conversely, dogs with lower attentional control or heightened anxiety may integrate new information more slowly, lowering η and resulting in more gradual learning curves. Thus, individual temperament shapes not just *whether* a dog learns well, but can also shed light on *how it updates* information over time, strengthening the computational lens in understanding variability in service-dog training outcomes.

Although our proposed framework only highlights the use of the RW and SARSA models, other learning models could also explain different types of task training and their stages (e.g. a trained dog learning to optimize application of its training to new tasks). Ultimately, empirical work applying computational learning models directly to service dog training is limited, but this section highlighted the potential of using formal models to optimize training protocols, individualize reinforcement strategies for training, and deepen our scientific understanding of how service animals learn.

Conclusion

Our findings surrounding various aspects of service dog training are valuable in that they address multiple constraints in the current approaches. Existing systems for training service dogs are costly, time consuming, and generally rely on dog breeding rather than the use of shelter dogs. Incorporating the neuroanatomical findings of Hecht et al. (2021) into service dog selection could be one future of this field. A system of training that looks at dogs on an individual basis, rather than by typical breed traits, will not only allow for a greater availability of service dogs sourced from shelters, but also address the fact that about half of dogs in training are failing out of the programs. Prescreening through anatomical MRI scans could help predict whether or not an individual dog may be well suited for service work based on their brain anatomy. While this potential solution addresses some pitfalls of the current training and selection system, it is not without its limitations. MRI screening for individual dogs is expensive and impractical right now, and a purely neuroanatomical approach to service dog selection ignores the observed behavioral history and traits of a given dog. This is where LCA could be integrated to account for the constraints that a neuroanatomy-only model imposes, by incorporating behavioral and experiential variability into predictions of training success. Our findings also show how different combinations of computational learning models can be used to map how service dogs are trained for various purposes. Each service dog requires specific training depending on the task or duty they need to perform. It is helpful to view training in different stages, as modeled by different learning models, to understand what types of learning take place during each process.

The field of service dog temperament and learning research is relatively new and small. The lack of research in this field is a major limitation for our findings as we have had to pull some of the existing studies about dogs in general to the use and training of service dogs. Further research in the field could benefit from studies that include groups of dogs that have succeeded in service dog training and those that have not, as well as dogs from a variety of backgrounds. Another limitation of this research is that there is no specific type of training for all service dogs. There is no single standard way to train a service dog, which makes sense given that different service dogs perform a variety of different tasks. Some dogs may be trained for item retrieval, whereas others may be trained for medical alert, two examples that show how wide the range of training can be. The lack of a standard training procedure is a barrier to research since finding and training a ‘good service dog’ is highly dependent on the task at hand and what behavioral factors contribute to a dog’s ability to perform the service task. Future studies would therefore benefit from examining particular service tasks individually, rather than treating all service roles as equivalent, which would better capture the distinct learning demands each task entails.

A future direction in this field of research could focus on designing a questionnaire for dogs that is not as time-and engagement-dependent as C-BARQ. Creating a questionnaire that can be reliably used for both shelter dogs and pure-bred dogs would increase the scalability of

screening, and in turn expand the pool of potential service dog candidates. Furthermore, although there is substantial work on the effectiveness of different reward types, additional progress could come from mapping these reward preferences onto computational learning models. Examining whether different reward types carry distinct subjective values for individual dogs would deepen our understanding of canine learning mechanisms and enable more targeted, dog-specific training strategies that account for internal motivational states.

Contributions:
Katie: Wrote introduction and latent class analysis paragraphs within the temperament section. Helped with clarifications in the revision process. Found 7 out of the 9 studies/resources used in this paper. Equally brainstormed during project duration. Heavily responsible for incorporation, comprehension, and presentation of computational topics (machine learning models, LCA, and step by step breakdown of SARSA application).

Annie: Wrote the Hecht et al. literature review and the section on the Rescorla–Wagner model. Analyzed the neural and neuroimaging papers we incorporated, taking responsibility for understanding how their findings fit into the broader aims of the project and for presenting this material to the class. Emphasized the importance of integrating computational learning models from class with temperament research to ensure that our framework was cohesive and not disjointed. Proposed examining how multiple learning models (rather than a single approach) might account for the complexity of service-dog training. In addition to writing my assigned sections, I was in charge of editing and refining my group members' sections, offering suggestions to improve clarity and flow, and identifying areas where key details needed to be added.

Aarushi: Wrote the learning models introduction and conclusion section, as well as the explanation for the application of the SARSA model. Worked with Annie to integrate multiple computational learning models into our explanation. Worked with Katie to conceptualize and apply the SARSA model to our task example, to bridge the gap in literature; and took point on presenting it clearly in the class presentation. Researched more of the behavioral components of the earlier parts of our research query (more in Presentation 1, did not end up making a big part of our final work, more as potential next steps). Equally brainstormed during project duration, and helped with clarifications and writing flow in the revision process. Took point on all presentation design and organizational structure (project documents etc).

Kalena: Wrote the conclusion section of the paper. Involved in editing of the conclusion for the final paper. Equally brainstormed during project duration and helped with outlining the paper. Involved in the literature search process; contributed to a list of potential papers to include in the project.

Honor Code Statement: *This paper represents my own work in accordance with University regulations.*

/s/ Anne Townsend, Katie Bausenwein, Aarushi Adlakha, Kalena Bing

References

Acebes, F., Pellitero, J. L., Muñiz-Diez, C., & Loy, I. (2022). Development of Desirable Behaviors in Dog-Assisted Interventions. *Animals : An Open Access Journal from MDPI*, *12*(4), 477. https://doi.org/10.3390/ani12040477

ADA Requirements: Service Animals. (2020). ADA.Gov. https://www.ada.gov/resources/service-animals-2010-requirements/

Cleghern, Z., Gruen, M., Roberts, D. (2018) Using decision tree learning as an interpretable model for predicting candidate guide dog success. *Measuring Behavior 2018*. Manchester: Manchester Metropolitan University, pp. 252-258

Duffy, D. L., & Serpell, J. A. (2012). Predictive validity of a method for evaluating temperament in young guide and service dogs. *Applied Animal Behaviour Science*, *138*(1), 99–109. https://doi.org/10.1016/j.applanim.2012.02.011

Fukuzawa, M., & Hayashi, N. (2013). Comparison of 3 different reinforcements of learning in dogs (canis familiaris). *Journal of Veterinary Behavior*, *8*(4), 221–224. https://doi.org/10.1016/j.jveb.2013.04.067

Harvey, N. D., Craigon, P. J., Blythe, S. A., England, G. C. W., & Asher, L. (2017). An evidence-based decision assistance model for predicting training outcome in juvenile guide dogs. *PLOS ONE*, *12*(6), e0174261. https://doi.org/10.1371/journal.pone.0174261

From puppy to partner: How service dogs are trained. UDS Foundation. (2025, November 6). https://udservices.org/from-puppy-to-partner-how-service-dogs-are-trained/

Service Dogs vs. ESD vs. Therapy Dog: What's the Difference? (n.d.). Retrieved December 5, 2025, from https://www.mncanineconsulting.com/post/service-dogs-esd-therapy-dog

Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent class analysis: A guide to best practice. Journal of Black Psychology, 46(4), 287–311. https://doi.org/10.1177/0095798420930932

Zapata, I., Eyre, A. W., Alvarez, C. E., & Serpell, J. A. (2022). Latent class analysis of behavior across dog breeds reveal underlying temperament profiles. *Scientific Reports*, *12*(1), 15627. https://doi.org/10.1038/s41598-022-20053-6