# Which personality dimensions do puppy tests measure? A systematic procedure for categorizing behavioral assays

**ARTICLE** *in* BEHAVIOURAL PROCESSES · SEPTEMBER 2014

Impact Factor: 1.57 · DOI: 10.1016/j.beproc.2014.09.029



**3 AUTHORS**, INCLUDING:

David L. Sinn

University of Tasmania

**26** PUBLICATIONS **781** CITATIONS

SEE PROFILE

Samuel D Gosling

University of Texas at Austin

**124** PUBLICATIONS **11,024** CITATIONS

SEE PROFILE

Behavioural Processes 110 (2015) 117–124

Contents lists available at ScienceDirect

Behavioural Processes

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

# Which personality dimensions do puppy tests measure? A systematic procedure for categorizing behavioral assays

Monica E. McGarrity[a,b], David L. Sinn[a,c,*], Samuel D. Gosling[a]

[a] University of Texas at Austin, Department of Psychology, 108 E. Dean Keeton Stop A8000, Austin, TX 78712-1043, USA
[b] Texas Parks and Wildlife Department, 4200 Smith School Rd., Austin TX 78744, USA
[c] University of Tasmania, School of Biological Sciences, Private Bag 5, Hobart, TAS 7001, Australia

ARTICLE INFO

*Article history:*
Available online 30 September 2014

*Keywords:*
Animal personality
Measurement science
Expert-categorization
Standardized test
Domestic dog

ABSTRACT

With the recent increase in interest in personality in dogs, behavioral assays of their behavior have proliferated. There has been particularly strong interest in predicting adult behavior from puppy tests. As a result, researchers and practitioners seeking to measure personality in puppies are faced with a bewildering array of options and no clear guide as to what behavioral assays have been developed or which personality dimensions those assays measure effectively. To address this issue, we used an 'expert-categorization' procedure—a standardized method often used in the course of meta-analyses—to identify the subset of those assays consensually judged to measure major personality dimensions effectively. We used this procedure to identify all relevant puppy tests and to categorize them in terms of their ability to measure nine personality dimensions identified in previous research (activity, aggressiveness, boldness/self-assuredness, exploration, fearfulness/nervousness, reactivity, sociability, submissiveness, trainability/responsiveness). Specifically, we identified 264 assay subtests, derived from 47 studies, which were subjected to a standardized categorization procedure undertaken independently by six expert judges. These procedures yielded a set of behavioral tests judged to measure the nine dimensions effectively and also demonstrated a widely applicable method for developing and evaluating behavioral test batteries.

This article is part of a Special Issue entitled: Canine Behavior.



## 1. Introduction

Recent decades have witnessed a strong and steady growth in research on personality or temperament in dogs (Fratkin et al., 2013; Jones and Gosling, 2005). This growth has been driven, in large part, by practical and applied questions about the extent to which the adult behavior of working or companion dogs can be predicted from their behavior as puppies (e.g., Bollen and Horowitz, 2008; Campbell, 1972; Duffy and Serpell, 2012; Goddard and Beilharz, 1985; Hackney 1986; King et al., 2012; Sinn et al., 2010; Svobodova et al., 2008; Wilsson and Sinn, 2012; Wilsson and Sundgren, 1998a). The behavioral tests (or "assays") have been developed across a wide array of theoretical, practical, and commercial contexts. Each of these contexts is associated with unique goals and perspectives on measurement. Consequently, there has been a proliferation of behavioral assays, designed to assess a wide array of personality traits.

The creation of so many tests reflects the healthy state of canine science. However, researchers and practitioners wanting to measure personality in puppies are faced with a bewildering array of options and no clear guide as to what behavioral assays have been developed or which personality dimensions those assays measure. To address this issue, we used a systematic procedure to identify all relevant puppy tests and categorize them in terms of their ability to measure different personality traits. In doing so, we also provide a demonstration of the procedure, which can be adapted for use in other contexts (e.g., identifying behavioral assays for special populations).

### 1.1. How many personality dimensions?

With so many researchers from so many different disciplines interested in dogs, there is little agreement on the number of dimensions needed to characterize the full range of individual

* Corresponding author at: University of Texas at Austin, Department of Psychology, 108 E. Dean Keeton Stop A8000, Austin, TX 78712-1043, USA.
*E-mail address:* david.sinn@utexas.edu (D.L. Sinn).

http://dx.doi.org/10.1016/j.beproc.2014.09.029
0376-6357/

**Table 1**
Personality dimension framework used as the basis for the judges' categorizations of the behavioral assay subtests.

| Source reviewed | Personality dimension | Representative descriptors for each personality dimension |
|---|---|---|
| Jones and Gosling (2005), Réale et al. (2007) and Ley et al. (2008)[a] | Activity | Locomotor activity, activity/inactivity, energy, hyperactivity, restlessness |
| Jones and Gosling (2005) and Réale et al. (2007) | Aggressiveness | Aggression, bite, agonistic reaction |
| Réale et al. (2007) and Ley et al. (2008) | Boldness/self-assuredness | Determination, tenaciousness, independence, opportunistic, tameness (tame/untamed), shyness (shy/not shy), behavior in response to any risky (but not new) situation |
| Réale et al. (2007) | Exploration | Exploratory, neophilia, curiosity, behavior in response to a new situation |
| Jones and Gosling (2005) and Ley et al. (2008) | Fearfulness/nervousness | General nervousness, flight, timidity, wariness, cautiousness, sensitivity |
| Jones and Gosling (2005) | Reactivity | Excitability, behavior problems |
| Jones and Gosling (2005), Réale et al. (2007) and Ley et al. (2008)[b] | Sociability | Affability, extraversion, playfulness, behavior in response to the presence or absence of conspecifics (excluding aggressive behavior) |
| Jones and Gosling (2005) | Submissiveness | Submission, dominance |
| Jones and Gosling (2005) and Ley et al. (2008) | Trainability/responsiveness | Obedience, distractibility, cooperation, reliability, trainability, intelligence, attentiveness, cleverness |
| N/A | Other dimension | Judge names other dimension not listed above |
| N/A | None of the above dimensions | Judge determines test does not seem to measure any of the above dimensions |
| N/A | Unable to determine | Judge unable to determine dimensions measured based upon information provided in the study description |

[a] Ley et al. (2008) *'extraversion' dimension descriptors overlap with 'activity'.*
[b] Ley et al. (2008) *'amicability' dimension descriptors overlap with 'sociability'.*

differences in dog personality. For example, one model of canine personality specifies three broad underlying dimensions (Svartberg, 2005), but others have suggested 5 (Svartberg and Forkman, 2002), 6 (Svartberg et al., 2005) or as many as 11 (Duffy and Serpell, 2012; Serpell and Hsu, 2001; Hsu and Serpell, 2003). To empirically evaluate the points of convergence and divergence across the models, one review of the dog-personality literature collated all previous work on dog-personality structure, subjecting the constituent behaviors and traits to an expert-categorization procedure; these procedures identified seven broad dimensions of personality: activity, aggressiveness, fearfulness/nervousness, reactivity, sociability, submissiveness, and trainability/responsiveness (Jones and Gosling, 2005). In addition, Réale et al. (2007) developed a five-dimension model that includes two dimensions—boldness/self-assuredness and exploration—not fully captured by the seven (Jones and Gosling, 2005) categories. Combining the seven dimensions identified by Jones and Gosling (2005) with the two additional dimensions identified by Réale et al. (2007) yields a framework (Table 1) that is both parsimonious and covers all the major dimensions of personality (e.g., Hsu and Serpell, 2003; Ley et al., 2008; Mirkó et al., 2012; Svartberg and Forkman, 2002).

### 1.2. Expert-categorization procedure

The research on dog personality also reveals considerable disagreement over terminology, with some researchers using similar terms to describe different behaviors and other researchers using different terms to describe similar behaviors (i.e., jingle/jangle fallacies: Carter et al., 2013; Gosling, 2001). Consequently, it is often unclear what the behavioral assays used in dog studies actually measure. To address this issue, a systematic procedure is needed for determining which personality dimensions are expressed in the various behavioral assays. One promising approach involves the use of the 'expert-categorization' procedures typically used in the course of quantitative meta-analyses (Barrick and Mount, 1991; Fratkin et al., 2013; Lipsey and Wilson, 1996; Rosenthal, 1991), in which subject-matter experts independently sort items into categories and their categorizations are subsequently aggregated statistically. The advantage of this approach is that it can guard against the idiosyncratic biases of any one individual that result from variation in type and breadth of experience among individuals. This method uses a panel of expert judges (i.e., individuals with substantial experience working with behavioral research and the species or taxon of interest) who independently categorize items according to a personality dimension framework. The items sorted are removed from their theoretical context (e.g., the assay purportedly measuring sociability is not labeled as such when it is categorized), allowing the expert judges to focus on the content whilst blind to the original designation of the item.

### 1.3. Review and categorization of behavioral assays

The aim of the present review is to identify the full range of behavioral assays used in previous research and demonstrate the use of an 'expert-categorization' procedure—often used in the course of meta-analyses—to identify the subset of those assays that measure major personality dimensions. In light of the applied angle of much of the research done on dogs, many canine-personality research studies focus on predicting adult behavior using puppy tests (Fratkin et al., 2013). Our goal was to identify assays that are applicable across a wide range of contexts. Therefore, our analysis focused on assays evaluating behavioral traits of puppies (i.e., dogs less than 12 months of age) for the purposes of characterizing personality.

This aim was achieved in two phases. The first phase accomplished a literature review and identified, for categorization purposes, the discrete behavior assay subtests described in each selected publication. The second phase demonstrated the use of standardized meta-analytic 'expert-categorization' methods to determine which personality dimensions are measured by behavioral assays. We used the results of our categorization procedures to provide insights regarding behavioral assays that are likely to be useful for measuring personality in future research.

## 2. Materials and methods

### 2.1. Phase 1: literature review

#### 2.1.1. Literature search procedures

We conducted a comprehensive literature database search in July 2012 using a matrix of search phrases generated using all possible combinations of the terms "puppy," "dog," and "canine," crossed with the terms "temperament," "personality," and "behavior." We used each combination of search terms and searched the Biosis, PsychInfo, Web of Science, and Google Scholar databases. We restricted Web of Science searches to the behavioral sciences and psychology categories and searched Google Scholar only for articles published since 2000, considering only the top 500 results when sorted by relevance. Including Google Scholar results earlier than 2000 did not yield any published articles that had not been identified through Biosis, PsychInfo, and Web of Science searches; we considered only the top 500 results in Google Scholar searches because after only 300 results no new articles were ever identified.

#### 2.1.2. Inclusion criteria

Prediction of adult behavior for applied purposes (e.g., determination of suitability for working roles) using a battery of standardized puppy tests is an important goal of many canine-personality studies (Fratkin et al., 2013). To address this need for batteries of standardized puppy tests that measure broad personality dimensions, we chose to focus on publications that described standardized behavioral assays to measure multiple personality traits in puppies (i.e., dogs up to 12 months of age: Andersen and Simpson, 1973; Morey, 1994). Therefore, we did not include studies that tested only adult dogs or used only questionnaire data rather than standardized tests (e.g., Batt et al., 2009; Goddard and Beilharz, 1983; Goodloe and Borchelt, 1998; Mirkó et al., 2012; Serpell and Hsu, 2001). Furthermore, our goal was to focus on broad characterizations of puppy personality so we did not include tests that focused on understanding of specific subsets of narrow behavior (e.g., agonistic behavior; Goddard and Beilharz, 1985). We found that some studies used puppy behavior tests only to evaluate training techniques (e.g., Stanley and Elliot, 1962) or evaluate canine response to a drug (e.g., Knowles et al., 1989), but did not seek to characterize puppy personality; the aim of our review was to identify standardized personality tests so we did not include these studies. Finally, we included only works written in English, as translation of non-English works was beyond the scope of this review and also likely beyond the capability of most researchers seeking to develop a behavioral assay.

#### 2.1.3. Literature search results

Our search procedures and inclusion decisions yielded 49 relevant publications – 38 journal articles, three master's theses, two doctoral dissertations, three books, one conference abstract/poster, one government report, and one magazine article published in the *American Kennel Gazette* (Bartlett, 1979). The temporal distribution of these 49 publications reflects the recent increase in research on this subject; more than half (i.e., 53%) were published between 2000 and 2012. The vast majority of these publications (90%) presented empirical, peer-reviewed research, but we also encountered one review article (Hackney 1986) and four technical articles that described personality assays but did not present empirical research (Bartlett, 1979; Campbell, 1972, 1975; Lindsay, 2005). Of these 49 publications, one book chapter (Campbell, 1975) reviewed a behavioral assay previously published by the same author (Campbell, 1972), neither of which presented empirical research; so for subsequent analyses, we considered these as one discrete study. Similarly, two journal articles reported research using the same behavioral assay and the same set of measurements for different analyses (Wilsson and Sundgren, 1998a, 1998b); we also considered these as one discrete study. Treatment of these publication pairs as non-discrete items yielded a final total of 47 works.

Campbell's (1972) Puppy Behavior Test (PBT) was the most commonly used behavioral assay (Batt et al., 2008; Beaudet et al., 1994; Campbell, 1972, 1975; Clark, 1994; Perez-Guisado et al., 2008; Thompson, 2005); no other assay was used more than twice. Population groups assayed included companion, shelter, guide/service, laboratory-reared, and police/guard/detection dogs; however, 38% of studies did not identify a population group. The breeds assayed represented 9 of the 10 major breed groupings of the Fédération Cynologique Internationale; however, the breed was unspecified or mixed/mongrel for 57% of studies. The four most commonly assayed breeds were German shepherd dogs, Labrador retrievers, beagles, and golden retrievers. Age groups assayed most frequently included 6, 7, 8, 11, and 12 weeks and 4, 6, or 9 months, with 8 weeks as the most prevalent testing age (49%).

#### 2.1.4. Identification of assay subtests

Most behavioral assays were named and scored as separate subtests (e.g., Campbell, 1972; "social attraction" test). However, this distinction between extremely different subtests was not always clearly made by the authors (i.e., no separate names or subheadings). For example, one study provided a narrative description of a test that involved a "fear walk" in a stimulating environment after which "we stopped in a quiet place and tested the dogs' reaction to 4 stimuli" (Goddard and Beilharz, 1984); these two portions of the assay were temporally separated and scored separately and were therefore described as separate subtests for categorization. This separation of functional subtests was essential for enabling the expert-categorization process to determine which portion of the test measured a dimension. Therefore, we created separate descriptions of 264 subtests defined by the authors of these 47 studies to facilitate and simplify the expert-categorization process; when subtests were not clearly identified, we separated subtests based on clear temporal and/or scoring distinctions. For each of the 264 subtests, we created a description of the test scenario (i.e., location, testers, stimuli, duration, etc.).

The measurements collected in conjunction with a behavioral assay are integral to determining what personality dimensions, or traits thereof, are measured by any behavioral assay. Therefore, we also compiled detailed descriptions of the scoring systems (e.g., codings such as frequency of a behavior; ratings of intensity of a behavior). The majority (62%) of publications used ratings scales to quantify behavior (e.g., Campbell, 1972). Nearly half of publications (47%) used behavioral codings such as frequency or latency (e.g., Dowling-Guyer et al., 2011). Several publications (8%) used ratings in conjunction with behavioral codings (e.g., ratings plus frequency or duration of behaviors; Wilsson and Sundgren, 1998a, 1998b). The level of detail of the description of scoring scales varied greatly; in many cases, researchers described only the anchors of ratings scales (e.g., 1 = "low amount" of behavior, 5 = "high amount" of behavior). To standardize the test descriptions as much as possible, we sought to create descriptions of each scale level based on the information provided (e.g., 2 = "greater amount" of behavior, 3 = "even greater amount" of behavior).

### 2.2. Phase 2: assay categorization

We selected a panel of six judges to demonstrate this procedure. Five of the six judges were current or former members of our lab group; however, the judges had diverse backgrounds that included behavioral experience with diverse taxa (e.g., academic research with canines, felines, other mammals, and invertebrates; commercial behavior research; working dog training; marine mammal

training; pet dog training; shelter dog assessment; See Supplementary Table S1) and all were familiar with the personality research literature in general. The diversity of the panel reduces the possibility that their aggregate judgments reflect only a narrow perspective on the behavioral assays and increases the likelihood that their aggregated judgments would generalize to other populations of judges (Block, 1961). Furthermore, each of our judges had considerable experience both in animal behavior research (mean = 13 years; median = 10 years; range = 4–30 years) and working with dogs in various capacities including, but not limited to, research (mean = 13 years; median = 10 years; range = 4–36 years). Therefore, we believe that, despite their diverse backgrounds, each judge was sufficiently familiar with dog behavior to make informed categorization judgments. Nonetheless, readers are cautioned to consider the size and constitution of the judge panel as they evaluate the consensus findings.

One of the judges was also a co-author of this review, however their input prior to survey completion was entirely conceptual and they were not privy to literature review results or survey details. Furthermore, this judge was found to be no more essential to attaining consensus than any other judge (i.e., there was little variation among judges in the degree to which they promoted proportion of inter-judge consensuses that would have been attained even without their "vote").

To facilitate categorization of subtests by dimension measured, we created an online survey describing the scenario and scoring of the 264 subtests. The categorization survey was hosted on the secure, web-based Research Electronic Data Capture application (REDCap; https://redcap.prc.utexas.edu). The judges were provided with a copy of a personality dimension framework with descriptors (Table 1) as a reference to use while completing the survey. The online survey presented judges with subtest descriptions one at a time in a randomized order and asked them to use checkboxes to indicate all of the nine dog personality dimension(s) each subtest seemed to "measure well." We used the term "measured well," not in regards to test validity but rather to encourage the judges to identify subtests that unambiguously measured a dimension effectively rather than a less stringent criterion of subtests that *may* or *might* measure that dimension. In addition to having the option of indicating multiple a priori dimensions, judges could indicate that a subtest measured 'none' of these dimensions "well," measured some 'other' dimension "well" (and name that 'other' dimension), or that they were 'unable to determine' which dimensions were measured "well."

The judge categorizations yielded measures of consensus and fidelity for each behavioral assay. Consensus reflects the degree to which the judges agreed that the assay measured one or more personality dimensions. Fidelity reflects the degree to which assays tap a narrow range of personality dimensions; the highest level of fidelity is 1, reflecting an assay that serves as a measure of one and only one personality dimension.

Following the dictionary definition of consensus (i.e., majority), whenever four or more of the six judges chose the same personality dimension for the same subtest, we considered consensus to have been reached that the dimension was measured by that subtest. In cases where judges identified an 'other' dimension that was a trait or descriptor listed in Table 1 or a thesaurus synonym thereof, we corrected the response to indicate that dimension if it was not also indicated. For example, when judges indicated that a test measured an 'other' dimension named 'distractibility'—which is listed in Table 1 as a descriptor of the 'trainability' dimension—we corrected the response to indicate that the dimension 'trainability' was measured, unless that dimension had already been selected. To retain independence of answers, the judges were asked not to discuss their answers, the survey, or the personality dimensions with others during the entirety of the categorization procedures.

**Table 2**
Judge consensus of temperament dimensions measured by rated subtests.

| Dimension | Proportion of subtests that measure dimension (≥4 judges indicate) |
|---|---|
| Activity | 0.14 ± 0.04 |
| Aggressiveness | 0.08 ± 0.03 |
| Boldness/self-assuredness | 0.11 ± 0.04 |
| Exploration | 0.11 ± 0.04 |
| Fearfulness/nervousness | 0.21 ± 0.05 |
| Reactivity | 0.09 ± 0.03 |
| Sociability | 0.26 ± 0.05 |
| Submissiveness | 0.04 ± 0.02 |
| Trainability/responsiveness | 0.03 ± 0.02 |
| *None of the above dimensions* | 0.02 ± 0.01 |
| *Unable to determine* | 0.05 ± 0.02 |

*Note*: The nine dimensions listed in the judges' rating framework are described in Table 1. Results here are shown as the proportion of all 264 subtests (i.e., #/264) followed by 95% confidence intervals.

Fidelity was determined by counting the number of dimensions that an assay was consensually judged to measure. Assays with low fidelity tapped several different dimensions and assays with high fidelity tapped few dimensions, ideally just 1.

## 3. Results

### 3.1. Assay categorization

Of the 264 subtests (assays), 181 met our consensus criterion, indicating that these assays were judged to measure at least one personality dimension. Of those, 81 of subtests had low fidelity (flagged by blue shading in Supplementary Table S2), measuring more than one dimension. One hundred subtests had high fidelity, with 81 of those measuring one of the nine focal dimensions (flagged by green shading in Supplementary Table S2) and 19 of them measuring a dimension other than the 9 focal dimensions (flagged by yellow shading in Supplementary Table S2).

Some subtests had very low fidelity, measuring as many as five different dimensions. For example, judge consensus found that nearly half of the subtests (8/18) of the Match-Up Behavioral Evaluation (Dowling-Guyer et al., 2011) measured multiple dimensions; this ambiguity is likely a consequence of the measurement method which involves coding occurrence of the same 38 behaviors or body postures during each of 18 subtests. For 83 of the subtests the judges failed to reach consensus on what those tests measured.

The consensus findings indicated that the dog personality dimensions measured by the greatest proportion of subtests (see Table 2) were 'sociability' (26% of subtests) and 'fearfulness/nervousness' (21%), followed—in order of decreasing proportion of subtests measuring the dimension—by 'activity' (14%), 'boldness/self-assuredness' (11%), 'exploration' (11%), 'reactivity' (9%), 'aggressiveness' (8%), 'submissiveness' (4%), and 'trainability/responsiveness' (3%).

As shown in Supplementary Table S2, the judges consensually indicated that they were unable to determine which dimensions were measured by 12 assay subtests, the vast majority of which (9/12) belonged to a personality test for shelter animals (Bollen and Horowitz, 2008).

#### 3.1.1. Assays consensually judged to measure sociability

For 57% of the studies we considered, judges agreed that there was at least one subtest that measured 'sociability' (Supplementary Table S2). Of the 69 subtests that measured 'sociability,' judges agreed unanimously on 21, 7 of which were found to measure only this dimension. These subtests use nearly identical methods, although measurement methods differ, and all involve presenting the puppy with an unfamiliar human who seeks to initiate play

**Table 3**
Behavioral assay subtests for which judge consensus was highest and measurement of the framework dimension indicated was independent of measurements of other dimensions. Other subtests that measure each dimension are given in Supplementary Table S2.

| Dimension | Assay source | Behavioral assay subtest name | Consensus level |
|---|---|---|---|
| Sociability | Åkerberg et al. (2012) | Contact test | 6 |
| | Hackney (1986) (CCI) | People orientation | 6 |
| | Hennessy et al. (2001) | Phase 2: reaction to person in arena | 6 |
| | Kim et al. (2010) | Friendly approach of stranger | 6 |
| | Murphree and Dykman (1965) | Reaction to friendly man | 6 |
| | Murphree et al. (1967) | Reaction to friendly man | 6 |
| | Wilsson and Sundgren (1998a, 1998b) | Part I: contact I | 6 |
| Fearfulness/nervousness | Kopechek (2010) | Novel item test | 6 |
| Activity | Hennessy et al. (2001) | Phase 1: initial reaction to being placed alone in arena | 5 |
| | Martinek (1973) | Habituation | 5 |
| | Murphree and Dykman (1965) | Brief exploratory activity | 5 |
| | Murphree (1973) | Brief exploratory activity | 5 |
| | Murphree et al. (1967) | Brief exploratory activity | 5 |
| | Riemer et al. (2011) | Temperament test – behavior on blanket | 5 |
| | Scott and Marston (1950) | Activity | 5 |
| Exploration | Pluijmakers et al. (2010) | Exposure effects on exploratory behavior | 6 |
| Boldness/self-assuredness | Svobodova et al. (2008) | Negotiating obstacles | 5 |
| Reactivity | Krauss (1976) | Fear of novel stimuli test | 5 |
| Aggressiveness | Sforzini et al. (2009) | Bowl removal by unknown person | 6 |
| Submissiveness | Wright (1980) | Bone-in-pen test | 4 |
| Trainability/responsiveness | Krauss (1976) | Person distraction test | |
| | Lindsay (2005) | Ball play | 5 |

or vocally encourages the puppy (see citations in Table 3 for more details).

#### 3.1.2. Assays consensually judged to measure fearfulness/nervousness

For 57% of the studies we considered, judges agreed that at least one subtest measured 'fearfulness/nervousness' (Supplementary Table S2). Of the 55 subtests that measured 'fearfulness/nervousness,' judges agreed unanimously on 9, only one of which was found to measure only this dimension. This subtest involves presenting a puppy with a novel item and measuring behaviors such as crouching, trembling, vocalization, and avoidance (see citation in Table 3 for more details).

#### 3.1.3. Assays consensually judged to measure activity

For 55% of the studies we considered, judges agreed that at least one subtest measured 'activity' (Supplementary Table S2). Of the 37 subtests that measured 'activity,' judges agreed unanimously on only one subtest, which was found to be ambiguous (i.e., also measured exploration). Five judges agreed on 21 subtests, 9 of which were found to measure only the activity dimension. These subtests use nearly identical methods that typically involve placing a puppy into a test arena with a grid painted on the floor and measuring activity as the number of line crossings and sniffing or by recording, and later measuring, the track walked by the puppy during the test period (see citations in Table 3 for more details).

#### 3.1.4. Assays consensually judged to measure exploration

For 43% of the studies we considered, judges agreed that at least one subtest measured 'exploration' (Supplementary Table S2). Of the 20 subtests that measured 'exploration,' judges agreed unanimously on 7of them, only one of which was found to measure only this dimension. This subtest involves placing the puppy in an arena that contains unfamiliar objects such as large balls, vacuums, or toys and coding the number of times the puppy approached one of these novel items (see citation in Table 3 for more details).

#### 3.1.5. Assays consensually judged to measure boldness/self-assuredness

For 38% of the studies we considered, judges agreed that at least one subtest measured 'boldness/self-assuredness' (Supplementary Table S2). Of the 30 subtests that measured 'boldness/self-assuredness,' judges did not agree unanimously on any. Five judges were able to agree on 13 subtests, only one of which was found to measure only this dimension. This subtest involves a walk over obstacles (e.g., stairs) during which the puppy is rated based on speed and hesitation (see citation in Table 3 for more details).

#### 3.1.6. Assays consensually judged to measure reactivity

For 32% of the studies we considered, judges agreed that at least one subtest measured 'reactivity' (Supplementary Table S2). Of the 23 subtests that measured 'reactivity,' judges did not agree unanimously on any. Five judges were able to agree on 4 subtests, only one of which was found to measure only this dimension. This subtest involves a walk during which puppies are presented with novel stimuli (e.g., person opening umbrella, live snake) and the number of startle responses was recorded (see citation in Table 3 for more details).

#### 3.1.7. Assays consensually judged to measure aggressiveness

For 23% of the studies we considered, judges agreed that at least one subtest measured 'aggressiveness' (Supplementary Table S2). Of the 21 subtests that measured 'aggressiveness,' judges agreed unanimously on only one, which was found to measure only this dimension. This subtest involves an unknown person who takes away the food bowl while the puppy is eating; behavior measurements include tail and ear position, eating speed, growling, and head raising (see citation in Table 3 for more details).

#### 3.1.8. Assays consensually judged to measure submissiveness

For 19% of the studies we considered, judges agreed that at least one subtest measured 'submissiveness' (Supplementary Table S2). However, for the 10 subtests that judges agreed measure 'submissiveness,' no more than four judges agreed on any subtest. Of these 10 subtests, only one was found to measure only this dimension.

This subtest involved placing all pairwise combinations of pups from a litter, two at a time, into a pen with a bone; observers recorded duration spent sharing or monopolizing possession of the bone (see citation in Table 3 for more details).

#### 3.1.9. Assays consensually judged to measure trainability/responsiveness

For 13% of the studies we considered, judges agreed that at least one subtest measured 'trainability/responsiveness' (Supplementary Table S2). Of the 9 subtests that measured 'trainability/responsiveness,' judges did not agree unanimously on any. Five judges agreed on three subtests, two of which were found to measure only this dimension. Subtests used to measure this dimension involve tossing a ball for a puppy and rating its tendency to chase and retrieve and its tendency to be distracted by (e.g., goes to person, notices person, ignores person) other people nearby (see citation in Table 3 for more details).

#### 3.1.10. Assays consensually judged to measure other dimensions

For 6% of the studies we considered, at least one judge indicated that at least one subtest measured one 'other' dimension, yet judges reached the minimum consensus level for four subtests measuring three judge-defined other dimensions—'toy/prey/hunt/chase drive' (two subtests; Krauss, 1976; Sinn et al., 2011), 'possession/resource guarding' (one subtest; Sforzini et al., 2009), and 'hunger/thirst/food motivation' (one subtest; Sforzini et al., 2009). For each subtest that a judge indicated measured some 'other' dimension, the same judge also indicated that the subtest measured one or more of the framework dimensions, suggesting that 'other' responses may represent additional, more specific traits/facets of these broad personality dimensions. Furthermore, 'hunger/thirst/food motivation' is not commonly considered a personality dimension, yet judges reached consensus for some subtests, agreeing that this "other" dimension was measured; the judges seemed to indicate that the subtest was measuring how hungry the puppy was rather than measuring any personality dimension. So, for these subtests, the dog's hunger or lack thereof could confound any other measurement the researcher was trying to make.

#### 3.1.11. Assays for which the judges appeared to be unable to identify a personality dimension

Notably, for all nine subtests of the Assess-A-Pet test (Bollen and Horowitz, 2008), judges agreed that they were unable to determine what dimension, if any, was measured; this was likely because the behavior was graded as 'pass/fail' and the authors provided little behavioral criteria upon which they based their original grade. Similarly, judges could not identify any dimension measured by two subtests used by Riemer et al. (2011) for which the authors provided little information on scoring. Lastly, judges were unable to determine any dimension measured by one subtest of the Puppy Temperament Test (Lindsay, 2005) in which the puppy is called by an evaluator on the other side of a curved wire mesh barrier—essentially a cognitive test—and is categorized into one of several subjective temperament types (e.g., 'sanguine') based on criteria that are not described.

## 4. Discussion

The present study sought to realize two goals. The first goal was to identify the behavioral assays in puppies consensually judged (by one group of experts) to measure the major dimensions of personality in dogs. The second goal was to demonstrate the categorization method in this context, with the hope that it can be applied in other contexts.

To achieve these goals, we first reviewed the literature to identify all relevant behavioral assays and then used a standardized categorization procedure to determine which personality dimensions those assays measured. Our analyses revealed that of the 264 behavioral assays, 81 measured one and only one of the focal dimensions (indicated by green shading in Supplementary Table S2), 81 others measured more than one dimensions (indicated by blue shading in Supplementary Table S2), and 102 measured none of the major dimensions.

The 'expert-categorization' approach differed from previous studies, which have categorized behavioral assays according to designation provided by the study authors or the originators of the measure (e.g., Jones and Gosling, 2005). Instead, we sought to reduce potential researcher biases by providing judges with standardized descriptions of the assays, without knowledge of the original researchers' intentions and by relying on the judges' consensus regarding which dimension(s) the assays measured. Nonetheless, it is important to recognize that even the judges making the categorizations in the present study will have brought their own idiosyncratic experiences to the categorization process and it is possible that another set of experts—or a larger set of experts—would have yielded different results. The background of our judges was diverse and the overlap in lab tenure was relatively brief (see Supplementary Table S1) so we believe that bias resulting from lab membership or a collegial relationship among our judges is unlikely to have inflated consensus estimates; moreover, research in the field of measurement shows that by aggregating the judgments of multiple experts the biasing effects of any experts' idiosyncrasies are greatly diminished (Block, 1961). Nonetheless, the possibility of bias generated by a restricted sample of judges does exist so we recommend that future studies using the expert-categorization technique recruit a larger cohort of judges from multiple research groups to minimize such potential bias.

We identified existing puppy behavioral assays that measure each of the nine broad personality dimensions defined in our framework effectively. Several behavioral assays covered in this review were found to measure multiple dimensions (see Supplementary Table S2); one assay (Krauss, 1976) was judged to measure all nine dimensions. Indeed, many assay subtests were found to have low fidelity, measuring multiple dimensions at once; typically such subtests were the ones that used scoring systems that recorded numerous facets of behavior from the same assay (e.g., Dowling-Guyer et al., 2011).

In some cases it may be desirable to measure dimensions separately (e.g., to evaluate the predictive potential or heritability of single dimensions). Therefore, researchers seeking to measure specific personality dimensions should take care to consider whether assay subtests and scoring systems measure only the desired dimension (those assay subtests with the green-shaded boxes in Supplementary Table S2). If suitable assays cannot be found that measure just a single dimension then the assays and scoring systems may require modification. Alternatively, statistical analysis can be applied to the multifaceted scoring systems to allow accurate differentiation of personality traits from the assays with lower fidelity. Based on our analyses, we identified personality-dimension specific insights as guidelines for researchers seeking to create their own battery of assays; however, researchers may wish to implement the 'expert-categorization' method demonstrated in this review to evaluate their own proposed battery of assays prior to implementation.

Furthermore, our report shows how the expert-categorization procedure can help to select a promising set of behavioral assays to measure dog personality but it does not guarantee that the assays are valid. For example, tests measuring food-related aggression (e.g., Marder et al., 2013) likely measure only one of many traits related to the 'aggressiveness' dimension and, if used alone,

may not have sufficient diagnostic value for making determinations regarding an animal's future (e.g., in a shelter context). Thus, it is important to keep in mind that our analyses reveal only which assays were consensually categorized by our judges to measure the dimensions and did so with high fidelity (i.e., measuring one and only one dimension) so they identify assays with high face validity but do not alone provide evidence for or against other forms of validity.

According to our judges' categorizations, 'sociability' can best be measured by presenting the puppy with an unfamiliar human and measuring (i.e., coding or rating) response to the human. 'Fearfulness/nervousness' can be measured by exposing the puppy to a novel item such as a device that makes a sudden, loud noise and measuring behaviors such as approach, avoidance, freezing, crouching, trembling, vocalization, or urination. 'Activity' can be measured by placing the puppy in an empty room and recording line crossings on a gridded floor. 'Exploration' can be measured by placing a puppy in an arena with novel objects (e.g., toys or vacuum) and measuring investigative behaviors such as approach of novel item. The lower fidelity of other subtests that measure 'exploration' can, in many cases, be attributed to overlap with measurement of the 'activity' dimensions (see Supplementary Table S2); therefore changes to the scoring systems could be targeted to reduce this overlap. 'Boldness/self-assuredness' can be measured by requiring the puppy to walk over obstacles and measuring speed and hesitation. 'Reactivity' can be measured by presenting a puppy with novel stimuli during a walk and measuring startle responses. 'Aggressiveness' may be measured by having a stranger remove the food bowl while the puppy is eating and measuring ear and tail position, eating speed, and growling; similar tests involve removing a rawhide treat (see Supplementary Table S2). Many other tests that measure aggressiveness have low fidelity, overlapping with measurements of multiple other dimensions. 'Submissiveness' may be measured by pairing puppies in a pen and recording sharing and dominant possession of a single bone. Other tests for 'submissiveness' with low fidelity involve measuring response (e.g., jumping, pawing, biting, growling, licking) to a person who crouches over and pets the puppy; however, these tests overlap with measurements of sociability and aggressiveness (see Supplementary Table S2). Finally, 'trainability/responsiveness' can be measured using a simple chase-retrieve sequence with a ball or other toy and scoring tendency to retrieve the toy; however, many of our judges indicated that this test also measures an 'other' dimension termed 'toy/prey/hunt/chase drive' that could arguably be one trait of the broader 'trainability/responsiveness' dimension, but could also potentially be a trait related to 'aggressiveness.' Therefore, independent measurement of the 'trainability/responsiveness' dimension may require development of a novel assay.

In summary, we hope our review and analyses serve to alert researchers to the array of tests available to measure personality in puppies. As a practical contribution to researchers and practitioners in applied fields, we have identified assays likely to be effective in measuring personality effectively. More broadly, we hope the expert-categorization procedures used here have demonstrated a method for resolving debate regarding the appropriate behavioral assays for a given dimension and will facilitate development of behavioral assays for future research.

## Acknowledgments

We thank M. Bensky, S. DeBono, J. Fratkin, A. Jones, D. Sinn, and S. Thomas for providing expert categorizations of personality subtests. This work was funded, in part, by the National Science Foundation Project (#CBET-0731216, NCE) and by U.S. Department of Homeland Security, Science & Technology Directorate research contract (HSHQDC-10-C-00085).

## Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.beproc.2014.09.029.

## References

Åkerberg, H., Wilsson, E., Sallander, M., Hedhammar, Å., Lagerstedt, A.-S., Larhammar, D., Meyerson, B., 2012. Test for personality characteristics in dogs used in research. J. Vet. Behav.: Clinic. Appl. Res.
Andersen, A.C., Simpson, M.E., 1973. The Ovary and Reproductive Cycle of the Dog (Beagle). Geron-X, Inc., Los Altos, CA.
Barrick, M.R., Mount, M.K., 1991. The big five personality dimensions and job performance: a meta-analysis. Pers. Psychol. 44, 1–25.
Bartlett, M., 1979. A novice looks at puppy testing. In: Purebred Dogs/American Kennel Gazette., pp. 31–42.
Batt, L.S., Batt, M.S., Baguley, J.A., McGreevy, P.D., 2008. Factors associated with success in guide dog training. J. Vet. Behav.: Clinic. Appl. Res. 3, 143–151.
Batt, L.S., Batt, M.S., Baguley, J.A., McGreevy, P.D., 2009. The value of puppy raisers' assessments of potential guide dogs' behavioral tendencies and ability to graduate. Anthrozoös 22, 71–76.
Beaudet, R., Chalifoux, A., Dallaire, A., 1994. Predictive value of activity level and behavioral evaluation on future dominance in puppies. Appl. Anim. Behav. Sci. 40, 273–284.
Block, J., 1961. The Q-sort method in personality assessment and psychiatric research. Thomas, Springfield, IL.
Bollen, K.S., Horowitz, J., 2008. Behavioural evaluation and demographic information in the assessment of aggressiveness in shelter dogs. Appl. Anim. Behav. Sci. 112, 120–135.
Campbell, W.E., 1972. A behavior test for puppy selection. Mod. Vet. Pract. 1972, 29–33.
Campbell, W., 1975. Problem Behavior in Dogs. American Veterinary Publications, Santa Barbara.
Carter, A.J., Feeney, W.E., Marshall, H.H., Cowlishaw, G., Heinsohn, R., 2013. Animal personality: what are behavioural ecologists measuring? Biol. Rev. 88, 465–475.
Clark, G.I., 1994. The Relationship between Emotionality and Temperament in Young Puppies. Colorado State University.
Dowling-Guyer, S., Marder, A., D'Arpino, S., 2011. Behavioral traits detected in shelter dogs by a behavior evaluation. Appl. Anim. Behav. Sci. 130, 107–114.
Duffy, D.L., Serpell, J.A., 2012. Predictive validity of a method for evaluating temperament in young guide and service dogs. Appl. Anim. Behav. Sci.
Fratkin, J.L., Sinn, D.L., Patall, E.A., Gosling, S.D., 2013. Personality consistency in dogs: a meta-analysis. PLOS ONE 8, e54907, http://dx.doi.org/10.1371/journal.pone.0054907.
Goddard, M.E., Beilharz, R.G., 1983. Genetics of traits which determine the suitability of dogs as guide-dogs for the blind. Appl. Anim. Ethol. 1982/83, 299–315.
Goddard, M.E., Beilharz, R.G., 1984. A factor analysis of fearfulness in potential guide dogs. Appl. Anim. Behav. Sci. 12, 253–265.
Goddard, M.E., Beilharz, R.G., 1985. Individual variation in agonistic behaviour in dogs. Anim. Behav. 33, 1338–1342.
Goodloe, L.P., Borchelt, P.L., 1998. Companion dog temperament traits. J. Appl. Anim. Welf. Sci. 1, 303–338.
Gosling, S.D., 2001. From mice to men: what can we learn about personality from animal research? Psychol. Bull. 137, 45–86.
Hackney, C.M., 1987. Guiding the Changes: Puppy Temperament Testing. University of Oklahoma.
Hennessy, M.B., Voith, V.L., Mazzei, S.J., Bruttram, J., Miller, D.D., Linden, F., 2001. Behavior and cortisol levels of dogs in a public animal shelter, and an exploration of the ability of these measures to predict problem behavior after adoption. Appl. Anim. Behav. Sci. 73, 217–233.
Hsu, Y., Serpell, J.A., 2003. Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. J. Am. Vet. Med. Assoc. 223, 1293–1300.
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.
Kim, Y.K., Lee, S.S., Il Oh, S., Kim, J.S., Suh, E.H., Houpt, K.A., Lee, H.C., Lee, H.J., Yeon, S.C., 2010. Behavioral reactivity of jindo dogs socialized at an early age compared with non-socialized dogs. J. Vet. Med. Sci. 72, 405–410.
King, T., Marston, L.C., Bennett, P.C., 2012. Breeding dogs for beauty and behaviour: why scientists need to do more to develop valid and reliable behaviour assessments for dogs kept as companions. Appl. Anim. Behav. Sci. 137, 1–12.
Knowles, P.A., Conner, R.L., Panksepp, J., 1989. Opiate effects on social behavior of juvenile dogs as a function of social deprivation. Pharmacol. Biochem. Behav. 33, 533–537.
Kopechek, M.E., 2010. Variation in the Onset and Expression of Hazard Avoidance Behavior across Three Breeds of Domestic Dogs. Ohio State University.
Krauss, J.L., 1976. The Predictive Value of a Puppy Test for Determining Future Trainability for Dog Obedience Work. Case Western Reserve University.

Ley, J., Bennett, P., Coleman, G., 2008. Personality dimensions that emerge in companion canines. Appl. Anim. Behav. Sci. 110, 305–317.
Lindsay, S.R., 2005. Appendix D: puppy temperament testing and evaluation. In: Handbook of Applied Dog Behavior and Training. Blackwell Publishing Professional, pp. 761–772.
Lipsey, M.W., Wilson, D.B., 1996. Practical Meta-Analysis. Sage Publications, Newbury Park, CA.
Marder, A.R., Shabelansky, A., Patronek, G.J., Dowling-Guyer, S., D'Arpino, S.S., 2013. Food-related aggression in shelter dogs: A comparison of behavior identified by a behavior evaluation in the shelter and owner reports after adoption. Appl. Anim. Behav. Sci. 148, 150–156.
Martinek, Z., 1973. The process of habituation as a test of interindividual (typological) differences in behavior of dogs. Acta Neurobiol. Exp. 33, 791–801.
Mirkó, E., Kubinyi, E., Gácsi, M., Miklósi, Á., 2012. Preliminary analysis of an adjective-based dog personality questionnaire developed to measure some aspects of personality in the domestic dog (*Canis familiaris*). Appl. Anim. Behav. Sci. 138, 88–98.
Morey, D.F., 1994. The early evolution of the domestic dog. Am. Sci. 82, 336–347.
Murphree, O.D., 1973. Inheritance of human aversion and inactivity in two strains of the pointer dog. Biol. Psychiatr. 7, 23–29.
Murphree, O.D., Dykman, R.A., 1965. Litter patterns in the offspring of nervous and stable dogs. I. Behavioral tests. J. Nerv. Ment. Dis. 141, 321–332.
Murphree, O.D., Dykman, R.A., Peters, J.E., 1967. Genetically-determined abnormal behavior in dogs. Cond. Reflex 2, 199–205.
Perez-Guisado, J., Munoz-Serrano, A., Lopez-Rodriguez, R., 2008. Evaluation of the Campbell test and the influence of age, sex, breed, and coat color on puppy behavioral responses. Revue Canadienne De Recherche Veterinaire 72, 269–277.
Pluijmakers, J.J.T.M., Appleby, D.L., Bradshaw, J.W.S., 2010. Exposure to video images between 3 and 5 weeks of age decreases neophobia in domestic dogs. Appl. Anim. Behav. Sci. 126, 51–58.
Réale, D., Reader, S.M., Sol, D., McDougall, P.T., Dingemanse, N.J., 2007. Integrating animal temperament within ecology and evolution. Biol. Rev. 2007, 291–318.
Riemer, S., Müller, C., Huber, L., Range, F., Kersting, E., Virányi, Z., 2011. Can early temperament tests predict behavioral tendencies in dog puppies? J. Vet. Behav.: Clin. Appl. Res. 6, 79.
Rosenthal, R., 1991. Meta-Analytic Procedures for Social Research. Sage Publications, Newbury Park, CA.
Scott, J.P., Marston, M.-V., 1950. Critical periods affecting the development of normal and mal-adjustive social behavior of puppies. Pedagog. Semin. J. Genet. Psychol. 77, 25–60.
Serpell, J.A., Hsu, Y., 2001. Development and validation of a novel method for evaluating behavior and temperament in guide dogs. Appl. Anim. Behav. Sci. 72, 347–364.
Sforzini, E., Michelazzi, M., Spada, E., Ricci, C., Carenzi, C., Milani, S., Luzi, F., Verga, M., 2009. Evaluation of young and adult dogs' reactivity. J. Vet. Behav.: Clin. Appl. Res. 4, 3–10.
Sinn, D.L., Gosling, S.D., Hilliard, S., 2010. Personality and performance in military working dogs: reliability and predictive validity of behavioral tests. Appl. Anim. Behav. Sci. 127, 51–65.
Sinn, D.L., Hixon, G., Gosling, S.D., 2011. Tasks 2.2, 2.3, and 3.2 combined report—exploratory factor analysis, internal validity of aggregate behavior scales, and test–retest correlations in the TSA-CBDC behavioral test data. In: US Department of Homeland Security, Science and Technology Directorate, Contract HSHQDC-10-C-00085: "Improving the Effectiveness of Detector-Dog Selection and Training through Measurement of Behavior and Temperament", p. 72.
Stanley, W.C., Elliot, O., 1962. Differential human handling as reinforcing events and as treatments influencing later social behavior in basenji puppies. Psychol. Rep. 10, 775–788.
Svartberg, K., 2005. A comparison of behaviour in test and in everyday life: evidence of three consistent boldness-related personality traits in dogs. Appl. Anim. Behav. Sci. 91, 103–128.
Svartberg, K., Forkman, B., 2002. Personality traits in the domestic dog (*Canis familiaris*). Appl. Anim. Behav. Sci. 79, 133–155.
Svartberg, K., Tapper, I., Temrin, H., Radesater, T., Thorman, S., 2005. Consistency of personality traits in dogs. Anim. Behav. 69, 283–291.
Svobodova, I., Vapenik, P., Pinc, L., Bartos, L., 2008. Testing German shepherd puppies to assess their chances of certification. Appl. Anim. Behav. Sci. 113, 139–149.
Thompson, R.S., 2005. Temperament in Domestic Dogs (*Canis familiaris*): analysis of a puppy assessment. University of Washington.
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.
Wilsson, E., Sundgren, P.E., 1998b. 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.
Wilsson, E., Sundgren, P.-E., 1998a. Effects of weight, litter size and parity of mother on the behaviour of the puppy and the adult dog. Appl. Anim. Behav. Sci. 56, 245–254.
Wright, J.C., 1980. The development of social structure during the primary socialization period in German shepherds. Dev. Psychobiol. 13, 17–24.