Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Applied Animal Behaviour Science 95 (2005) 1–53
APPLIED ANIMAL BEHAVIOUR SCIENCE
www.elsevier.com/locate/applanim

Review

Temperament and personality in dogs (*Canis familiaris*): A review and evaluation of past research

Amanda C. Jones *, Samuel D. Gosling

The University of Texas at Austin, Department of Psychology, 1 University Station A8000, Austin, TX 78712-0187, USA

Accepted 4 April 2005
Available online 9 June 2005

Abstract

Spurred by theoretical and applied goals, the study of dog temperament has begun to garner considerable research attention. The researchers studying temperament in dogs come from varied backgrounds, bringing with them diverse perspectives, and publishing in a broad range of journals. This paper reviews and evaluates the disparate work on canine temperament. We begin by summarizing general trends in research on canine temperament. To identify specific patterns, we propose several frameworks for organizing the literature based on the methods of assessment, the breeds examined, the purpose of the studies, the age at which the dogs were tested, the breeding and rearing environment, and the sexual status of the dogs. Next, an expert-sorting study shows that the enormous number of temperament traits examined can be usefully classified into seven broad dimensions. Meta-analyses of the findings pertaining to inter-rater agreement, test–retest reliability, internal consistency, and convergent validity generally support the reliability and validity of canine temperament tests but more studies are needed to support these preliminary findings. Studies examining discriminant validity are needed, as preliminary findings on discriminant validity are mixed. We close by drawing 18 conclusions about the field, identifying the major theoretical and empirical questions that remain to be addressed.

Keywords: Personality; Temperament; Dog; *Canis familiaris*

* Corresponding author. Tel.: +1 512 471 0691; fax: +1 512 471 5935.
E-mail address: acjones@mail.utexas.edu (A.C. Jones).

0168-1591/$ – see front matter
doi:10.1016/j.applanim.2005.04.008

Contents

1. Introduction

Early in the 20th century, Nobel laureate Ivan Pavlov began a research program designed to identify the basic types of canine temperament (e.g., Pavlov, 1906). Despite this auspicious start, the study of temperament and personality in animals did not evolve into a major area of research except, of course, in humans. Yet, pet owners and practitioners working with dogs have long recognized that temperament is important. It influences an individual's behavior and responses to the environment. Groups interested in temperament have ranged from private dog owners and dogs breeders to professional animal handlers and animal-research scientists; they have been consumed with such practical issues as matching dogs to appropriate homes and with understanding basic theoretical issues in animal behavior. This research has striven to fulfill many goals, from identifying a puppy test that will predict adult guide-dog behavior (e.g., Goddard and Beilharz, 1984a,b, 1986), to examining the heritability of temperament traits (e.g.,

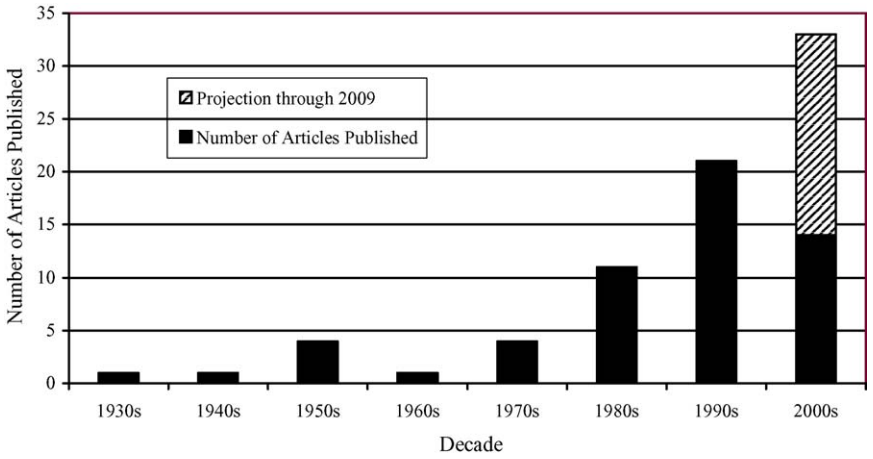

Fig. 1. Articles on dog temperament published each decade. Projection through 2009 is based on the number of articles published to date (i.e., in the 4 years and 3 months since January 2000).

Reuterwall and Ryman, 1973; Wilsson and Sundgren, 1998; Saetre et al., in press). Driven by such concerns, there has been a recent surge of published research on dog temperament (see Fig. 1).

This research is unified by a common interest in dog temperament, but the researchers conducting these studies come from a wide variety of backgrounds, bringing with them assorted perspectives and publishing in a broad range of journals. As a consequence of their distinct disciplinary affiliations and research goals, these efforts at understanding temperament in dogs have followed largely independent paths. The result is that it is hard to keep track of the various findings—the studies are scattered across journals in anthrozoology, psychology, biology, animal behavior, and veterinary medicine, among others.

Each of these discipline-bound studies is interesting and valuable in its own right, but it provides only a relatively narrow glimpse of dog temperament. Furthermore, there has been little effort to review and summarize what these numerous studies have taught us. The goal of this paper is to undertake such a review. This will bring together, for researchers in all disciplines, the diverse and disparate work on temperament in dogs. By doing so, we aim to identify general patterns of research and summarize the major findings to date. The review will also allow us to pinpoint the major gaps in our knowledge and determine what research challenges lay ahead.

Specifically, this review will start by examining general trends in research on dog temperament. What methods have been used, what breeds have been assessed, and what other trends can be identified? Next, we will review the studies of specific domains of temperament, identifying the temperament domains for which there is considerable cross-study support. Next, we will use meta-analyses of past work on the reliability and validity of temperament tests to evaluate the effectiveness of temperament measures. Finally, we will draw the findings together to offer 18 broad conclusions about the field and identify the major questions that remain to be addressed.

This review should be of interest both to practitioners and to researchers. Relevant practitioners include those interested in the practical task of identifying dogs temperamentally suited to working as guide dogs, hearing dogs or police dogs, and for various other jobs in which dogs assist people in their daily lives. The findings will also be relevant to dog shelters and rescue centers, which often rely on temperament tests as a guide for placing dogs in suitable homes, and for individual pet owners interested in finding a pet suitable for their lifestyle (e.g., Coren, 1995, 1998; Hart and Hart, 1985; Hart, 1995; Tortora, 1983). With the recent moves in the United States to pass breed-specific legislation, intended to limit and control the ownership of specific breeds, this work will also be of interest to workers in animal welfare and social policy. Finally, the review will be useful to the growing body of research scientists interested in using animal models to examine basic issues in human psychology (Gosling, 2001) and animal behavior (Dugatkin, 2004).

2. Definitions of temperament and personality

Before we begin a review of the temperament and personality literature, we must first determine what is meant by these terms and what, if any, difference exists between them. One seemingly trivial, yet pervasive, basis for distinguishing between temperament and personality is the disciplinary affiliation of the researchers associated with each term. Research on animals and human infants has tended to use the term *temperament* and research on human children and adults has tended to use the term *personality*. However, this distinction is not maintained consistently and the terms are often used interchangeably (McCrae et al., 2000).

In the human domain, temperament has been defined by some researchers as the inherited, early appearing tendencies that continue throughout life and serve as the foundation for personality (Buss, 1995; Goldsmith et al., 1987). Although this definition is not adopted uniformly by human researchers (McCrae et al., 2000), animal researchers agree even less about how to define temperament (Gosling, 2001). In some cases, the word ''temperament'' appears to be used purely to avoid using the word ''personality,'' which some animal researchers associate with anthropomorphism.

Most theoretical and empirical research on personality has been done in the human domain. Human-personality psychologists come in a variety of orientations and often differ in the personality constructs they emphasize. The phenomena studied by personality psychologists include temperament and character traits, dispositions, goals, personal projects, abilities, attitudes, physical and bodily states, moods, and life stories (John and Gosling, 2000). Thus, there is not one definition of personality that would satisfy all personality psychologists. Only a very broad (and thus somewhat vague) definition would satisfy most. For example, personality can be defined as those characteristics of individuals that describe and account for consistent patterns of feeling, thinking, and behaving (Pervin and John, 1997), a definition broad enough to capture most phenomena studied by personality psychologists.

Thus, the distinction between temperament and personality has not been maintained consistently in the literature. Given our goal to evaluate all potentially relevant studies, for

the purpose of this review we adopt a broad working definition that encompasses both constructs. In writing the article, our rule of thumb is to use the term ''temperament'' wherever possible but we also use the term ''personality'' where it is more appropriate to do so (e.g., when referring to work that explicitly discusses personality).

3. Literature review

This is the first major review of the field, so we decided to cast a broad net. Therefore, to be certain that our review uncovered as many potentially relevant studies as possible, we searched for all articles in the PsychInfo, Biosis, and Web of Science databases that examined either personality or temperament in dogs.

It is important to emphasize that our review included only those studies in the published empirical research literature. As such it did not include the methods that are frequently used and well-regarded in applied settings (e.g., Sue Sternberg's Assess-a-Pet and Emily Weiss's the SAFER test) but for which data are not yet publicly available.

3.1. Literature search procedures

Our literature search used two basic procedures: generating a large pool of potentially relevant articles, and selecting a smaller subset of articles for inclusion in the final review. These two procedures were used iteratively, such that one cycle generated a pool of potential articles and selected a subset of them for review, and this subset of articles provided the starting point for a subsequent cycle.

In the initial search cycle, we conducted searches in the PsychInfo, Biosis, and Web of Science databases for all articles containing the keywords ''dog'' and ''temperament,'' or ''dog'' and ''personality.'' We did not search for articles containing specific temperament descriptors such as ''aggressive'' or ''fearful'' because almost all behavior can be described as related to some domain. To include all such research would have cast our net too broadly, capturing a vast number of articles that were not really focused on personality constructs but had merely included behaviors related to a temperament domain. For example, the study of dogs' preference for humans by Topál et al. (1998) examined attachment behavior, including Nervousness-related behaviors, but had no interest in individual differences in temperament per se. Thus, we reasoned that if an article did not even mention personality or temperament in the title, list of keywords, or abstract (i.e., the fields scanned in a keyword search), it was highly unlikely that the research would be relevant to this review. The initial search cycle yielded 43 references from PsychInfo, 58 from Biosis, and 116 from the Web of Science.

After eliminating duplicates, we examined the abstracts of the remaining reports to eliminate irrelevant articles. Articles varied in their relevance to research on dog temperament; some focused directly on temperament assessment but others clearly fell beyond the domain of this review. For example, one article examined the personalities of people who strongly dislike dogs (Stubbs and Cook, 1999), and could therefore easily be classified as irrelevant. Although most articles could be unambiguously classified as clearly relevant or clearly irrelevant, there were a number of borderline reports that were distantly

or obliquely related to temperament but did not fall neatly into the core set of clearly relevant papers. We retained these borderline articles for closer inspection.

This review cannot include every vaguely relevant reference so only the most important borderline studies were retained. Given the goals of our review, we selected those articles that were empirical that were consistent with the definitions of temperament and personality described above, and that had a substantial focus on temperament or personality in dogs. Studies with only a cursory link to temperament were excluded. For example, we did not retain an article that described the working requirements for an animal-assisted therapy dog (Hart, 2000); it explained the functional significance and role of the therapy dog, touching only briefly on the temperament requirements.

Inspection of the references cited in the selected articles revealed several studies that had not been identified in the initial search. Therefore, each time a new article was identified, we searched its references for other relevant articles. After repeating this process several times, our leads began to run dry and we were satisfied that we had captured the vast majority of relevant research. Nonetheless, given the great diversity of research, we wanted to make sure our own disciplinary perspective did not bias the review. Therefore, we asked colleagues in other fields and who study dog behavior to check the reference list and bring to our attention any studies we had missed. By the end of these search procedures, we had identified 51 articles, all but one of which is summarized in Table 1. This study (Campbell, 1972) was retained because it is frequently referenced by and discussed in other studies, and because it seems to mark the beginning of a revival of interest in dog temperament.

Of course although we took care to identify all relevant articles, no selection procedure is flawless and we acknowledge that a few relevant studies will inevitably have slipped through our net. Nonetheless, we believe our review represents the most comprehensive summary to date of research on temperament and personality in dogs.

4. A general survey of the field

When reviewing a new field, the first major task is to step back and survey the general state of the field and identify the major trends. To this end, Table 1 summarizes the basic features of the studies included in our review. The first thing to note is that with one exception, there is a great diversity of research. The one exception comes in terms of the constructs studied; as in Gosling (2001) review of temperament in all non-human species, almost all the canine research has been on temperament traits, with almost no research on goals, motives, and other constructs.

In other respects, the studies are tremendously varied. They are drawn from a wide variety of disciplines, including animal behavior, biology, psychology, animal welfare, and veterinary medicine. The studies also have many different purposes, ranging from assessing temperament in specific breeds (e.g., Reuterwall and Ryman, 1973), to evaluating the domestic dog as a more general model of animal personality (e.g., Svartberg and Forkman, 2002). To help identify some specific patterns in this fragmented field, we propose several ways of summarizing the literature. These summaries are based on the methods of assessment, the breeds examined, the purpose of the studies, the age at which the dogs were tested, the breeding and rearing environment, and the sexual status of the animals.

Table 1
Summary of empirical research on dog temperament: study design, breed, sex, and age composition, and purpose of assessment

Study	*N*	Breed composition					Sex		Age at assessment (months)					Purpose of assessment					Population of Dogs
		GSD	Lab	Pure	Mixed	Unk.	M (neut)	F (spay)	First	Second	Third	Fourth	Fifth	Guide	Police	Work	Pet	Other	
Test Batteries																			
Cattell et al. (1973) and Cattell and Korth (1973)	101	0	0	101	0	0	NR (NR)	NR (NR)	.23–3.72	12				0	0	0	0	101	Research
Mahut (1958)	230	11	0	230	0	0	96 (NR)	134 (NR)	7–120					0	0	0	0	230	202 Privately owned, 20 show dogs, 8 research
Netto and Planta (1997)	112	NR	Yes	112	0	0	59 (NR)	53 (NR)	NR					0	0	0	112	0	Privately owned
Reuterwall and Ryman (1973)	958	958	0	958	0	0	NR[a] (NR)	NR[a] (NR)	18					Yes	Yes	Yes	0	0	Working dogs[b]
Royce (1955)	53	0	0	53	0	0	20 (NR)	33 (NR)	NR					0	0	0	0	53	Research
Ruefenacht et al. (2002)	3497	3497	0	3497	0	0	1679 (NR)	1818 (NR)	21.5[c]					0	0	0	0	3497[d]	Privately owned
Seksel et al. (1999)	60	NR	NR	50	10	0	32 (NR)	28 (NR)	1.38–3.91	+.46[e]	+.92[e]	+4 to 6[e]		0	0	0	60	0	Privately owned
Slabbert and Odendaal (1999)	167	167	0	167	0	0	NR (0)	NR (0)	1.85	2.77	3.70	6	9	0	167	0	0	0	Police work
Svartberg and Forkman (2002) and Saetre et al. (in press)	15329[f]	NR	NR	15329	0	0	7878[g] (NR)	7451[g] (NR)	19.72[c,h]					0	0	0	0	15329	Privately owned
Svartberg (2002)	2655	2219	0	2655	0	0	1381 (NR)	1274 (NR)	12–18					0	0	2655	0	0	Privately owned
van der Borg et al. (1991)	81	NR	NR	NR	NR	81	NR (NR)	NR (NR)	NR					0	0	0	81	0	Shelter dogs, adopted
Wilsson and Sundgren (1998)	630[s]	630	0	630	0	0	320 (0)	310 (0)	1.84	14.8–19.74				0	0	0	0	630	Police work, guide work, working dogs
Wilsson and Sundgren (1997)	2107	1310	797	2107	0	0	1073 (0)	1034 (0)	14.8–19.74					797	1310	0	0	0	Work/service, breeding
Total	25980	8792	797	25889	10	81	12538 (0)	12135 (0)	16.33[i], 18.94[j]					797	1477	2655	253	19840	
Ratings of individual dogs																			
Goodloe and Borchelt (1998)	2018	NR	NR	1412	588	18	916 [k] (613)	1084[k] (896)	NR					0	0	0	2018	0	Privately owned, show dogs
Gosling and Bonnenburg (1998)	1022	NR	NR	NR	NR	NR	NR (NR)	NR (NR)	NR					0	0	0	0	1022	Privately owned
Hsu and Serpell (2003)	2054[l]	48	94	1806	173	75	998 (NR[m])	1047 (NR[m])	62.20[c]					0	0	0	2054	0	Privately owned, show dogs
Ledger (2003)	234	15	0	234	0	0	NR (NR)	NR (NR)	NR					0	0	0	234	0	Privately owned
Podberscek and Serpell (1996)	1109	0	0	1109	0	0	545 (94)	564 (187)	3–204, mean = 32.4					0	0	0	1109	0	Privately owned, show dogs
Serpell (1983)	25	NR	NR	NR	NR	NR	NR (NR)	NR (NR)	NR					0	0	0	25	0	Privately owned
Serpell and Hsu (2001)	1067[n]	293	369	926	140	0	NR (NR[o])	NR (NR[o])	6	12	14–24			1067	0	0	0	0	Guide work
Stephen et al. (2001)	14	NR	NR	NR	NR	NR	NR (NR)	NR (NR)	NR					0	0	0	14	0	Privately owned

Table 1 (*Continued*)

Study	*N*	Breed composition					Sex		Age at assessment (months)					Purpose of assessment					Population of Dogs
		GSD	Lab	Pure	Mixed	Unk.	M (neut)	F (spay)	First	Second	Third	Fourth	Fifth	Guide	Police	Work	Pet	Other	
Wahlgren and Lester (2003)	264	<10[p]	37	216	48	0	119 (NR)	145 (NR)	NR					0	0	0	264	0	Privately owned
Total	7807	366	500	5703	949	93	2578 (707)	2840 (1083)	33.53[i], 40.21[j]					1067	0	0	5718	1022	
Expert ratings of breed prototypes[q]																			
Bradshaw and Goodwin (1998)	49	1	1	49	0	0								0	0	0	49	0	
Coren (1995)	79	1	1	79	0	0								0	0	0	79	0	
Draper (1995) [r]	56	1	1	56	0	0								0	0	0	0	56	
Hart and Miller (1985), Hart and Hart (1985), Hart (1995), and Hart et al. (1983)	56	1	1	56	0	0								0	0	0	56	0	
Keeler (1947)	5	0	0	5	0	0								0	0	0	0	5	
Lester (1983)	24	NR	NR	24	0	0								0	0	0	0	24	
Total	213	3	3	213	0	0								0	0	0	184	85	
Observational tests																			
Goddard and Beilharz (1984a,b, 1985)	102[s]	16	16	64	Yes	0	51 (51[t])	51 (0)	2.77	4	6	12	12–18	102	0	0	0	0	Guide dogs
Goddard and Beilharz (1982–1983)	887[s]	0	731	NR	NR	76	436 (227)	451 (0)	12–18					887	0	0	0	0	Guide dogs
Humphrey (1934)	NR	NR	0	NR	0	0	NR (NR)	NR (NR)	NR					0	0	Yes	0	0	Working dogs
James (1951)	11	0	0	11	0	0	5 (0)	6 (0)	NR					0	0	0	0	11	Research
Murphy (1998, 1995)	89	0	84	84	5	0	38 (NR[u])	51 (NR[u])	12[u]					89	0	0	0	0	Guide dogs
Total	1089	16	831	159	5	76	530 (278)	559 (0)	9.92[i], 13.60[j]					1078	0	Yes	0	11	
Studies that used a combination of methods																			
Beaudet et al. (1994)	39	15	0	0	0	0	15 (0)	24 (0)	1.61	3.68				0	0	0	0	39	Privately owned
Goddard and Beilharz (1986)	102[s]	16	16	64	Yes	0	51 (51[t])	51 (0)	.92[v]	1.15	1.38	1.62	1.85	102	0	0	0	0	Guide dogs
Gosling et al. (2003a)	78	NR	NR	NR	NR	NR	39 (NR)	39 (NR)	NR					0	0	0	0	78	Privately owned
Hennessy et al. (2001)	166	NR	NR	NR	NR	NR	70 (NR)	96 (NR)	NR[w]	+.46[e]	+6[e]			0	0	0	166	0	Shelter dogs
Ledger et al. (1995)	120	NR	NR	NR	NR	NR	NR (NR)	NR (NR)	NR					0	0	0	120	0	Shelter dogs
Ledger and Baxter (1996, 1997)	56[x]	NR	NR	NR	NR	NR	NR (NR)	NR (NR)	NR					0	0	0	56	0	Shelter dogs
Stephen and Ledger (2003)	40	NR	NR	NR	NR	NR	NR (NR)	NR (NR)	NR					0	0	0	40	0	Shelter dogs
Weiss and Greenberg (1997)	9	0	0	0	9	0	6 (NR)	3 (NR)	10–24					9	0	0	0	0	Shelter dogs

Total	610	31	16	64	9	0	181 (51)	213 (0)	6.51[i], 2.06[j]	111	0	0	382	117		
Grand total	35699[y]	9208	21	4732	028	973	250	15827 (1036)	15747 (1083)	16.48[i], 21.56[j]	3053	1477	2655	6537	21075	

Notes: N, the number of subjects in each study; GSD, German Shepherd Dog; Lab, Labrador Retriever; Pure, dogs of specific, unmixed breeds including GSDs and Labs; Mixed, dogs known to be of mixed breeding; Unk., dogs' breeds were unknown or not recorded, and guesses about mixed breeds were not made; M, male dogs; and Neut, neutered; F, female dogs; Spay, the number spayed. Age at assessment has the sub-groupings of first, second, third, fourth, and fifth because dogs may be tested more than once, at different ages. Guide indicates that these dogs were assessed for possible use as guide dogs. Police indicates that these dogs were assessed for possible use as police dogs. Work indicates that these dogs were assessed for possible use in other types of work (e.g., field work, search and rescue, tracking, protection work). Pet indicates that these puppies or dogs were assessed for selection as a pet, or that they already were pets at the time of assessment. The dogs in the category Other do not fit into any of the previous categories; they may be in studies seeking to learn more about personality itself. NR indicates that the authors did not report that particular piece of information, whereas "yes" indicates that the authors reported that there were in fact dogs of that type involved but did not report a number or percentage. When there is no entry, that calculation or report of the particular statistic is not appropriate or not applicable for the given study. One Test Batteries in our review (Campbell, 1972) is excluded from this table because it includes descriptions of how to test dog temperament, not actual evaluations; other articles in our review (e.g., Keeler, 1947; Roll and Unshelm, 1997) are not included because, though they discuss dog temperament, they do not present assessments of temperament.

[a] The number of male and female dogs in this study varies between analyses.

[b] The goal of Reuterwall and Ryman's (1973) article was to study the genetic components of behavior in German Shepherd Dogs; the test used was the Army Dog Training Center test which was designed to identify dogs suitable as working dogs and potentially to breed future generations of working dogs.

[c] These studies reported the average age of dogs assessed.

[d] Ruefenacht et al. (2002) gathered data through the Swedish German Shepherd Dog breeding club, which strives to evaluate whether dogs are physically and temperamentally sound enough for future breeding for many purposes (police work, guide work, protection work, etc.).

[e] The ages at each subsequent testing are reported in terms of number of days or months since the first testing.

[f] Five Dachshunds and five Sight hounds were excluded from the analyses because their breed groups were under-represented in the sample.

[g] The numbers of males and females in these studies are calculated using the percentages given by the authors, and are then rounded to the nearest whole number.

[h] All dogs were at least 12 months old when tested.

[i] The averages are calculated by adding together all the ages in one particular category (e.g. Test Batteries) and then dividing by the number of studies that reported age information; thus, studies that did not report an *N* do not skew (shrink) the average age. When the age in a given study is reported as a range (e.g. Cattell et al., 1973), the midpoint is used in calculating the overall average.

[j] Weighted average, weighted by number of dogs in each study.

[k] The authors note that the sex of 18 dogs is missing from the surveys they collected.

[l] The authors report 2054 dogs total, but also report 998 males and 1047 females, for a total of 2045 dogs.

[m] Hsu and Serpell (2003) report that 59% of the dogs in their study are neutered (castrated) but do not report how many of these dogs are male and female.

[n] The authors report the total number of dogs in this study is reported inconsistently as both 1067 and 1097 without explaining the discrepancy. We are reporting an *N* of 1067 because this is closer to the sum of the authors' report of dogs when broken down by breed (1066).

[o] All but 10 dogs are intact.

[p] The authors specify that there are >10 GSDs included in the study but do not give a precise number; they have not been included in the totals.

[q] The numbers in this section represent the breeds evaluated; no actual dogs were involved in the studies.

[r] This study is a reanalysis of the data collected by Hart et al. (1983); the 56 dogs in that study are included only once in the totals.

[s] The authors report that not all original subjects were maintained throughout the study, but do not indicate how many subjects were maintained. Where applicable, the number of dogs per breed is thus also uncertain, because we do not know the breed of individuals who dropped out.

[t] All male dogs were castrated at approximately six months of age.

[u] All but four dogs were castrated; those four dogs were ex-show or ex-breeding dogs, were donated to the guide dog program, and were several years older than the other dogs assessed.

[v] Assessments were conducted weekly until the puppy reached 6 months of age, then another was performed at 12 months.

[w] Hennessy et al. (2001) include dogs of varying ages, divided into two groups: "puppies," who still have milk teeth, and "juveniles/adults," who have their adult teeth. Ages are not reported.

[x] 56 dogs were originally tested, but follow-up surveys assessed only 40 of the original 56.

[y] The 102 dogs from the Goddard and Beilharz studies are counted towards the total number of dogs each time there is a separate listing for them, because different tests at different ages are analyzed.

4.1. Assessment methods

Table 1 is organized in terms of the four main methods by which dog temperament has been assessed: Test Batteries, Ratings of Individual Dogs, Expert Ratings of Breed Prototypes, and Observational Tests. A fifth category was composed of studies that combined more than one assessment method.

As shown in the table, the most common method of assessment was the Test Battery, which appeared as the primary assessment method in 33% of the 51 studies reviewed. The core goal of studies using this method was to document dogs' reactions to specific stimuli. The tests were performed by presenting various, usually novel, stimuli one at a time to a canine subject and recording its reaction(s). Thus, Test Batteries had two components: the tests themselves and the system for coding the dogs' reactions to the tests.

In theory, Test Batteries were the closest of the four methods to achieving objectivity, but in practice the levels of objectivity actually attained varied substantially. One of the more objective Test Battery studies examined the relationship between Fearfulness and breed (Mahut, 1958). After presenting novel stimuli to target dogs, the researchers described the dogs' subsequent behaviors purely in terms of what was visually and auditorily observed over the next 10 s.

The second most common method of assessment was Ratings of Individual Dog, appearing in 18% of the studies reviewed. The goal of these studies was to gather information about individual dogs' behaviors and histories from an informant. One such data-gathering technique was to have a particular dog's owner state whether or not, or how often, his or her dog had engaged in a specified behavior (e.g., snapping at children). The owners who participated in such studies were usually preselected on the basis of group membership (e.g., owners of a specific dog breed). For example, Podberscek and Serpell (1996) asked English Cocker Spaniel (ECS) owners how likely, on a five-point scale, their ECS was to act aggressively towards strange dogs, when reached for by a person, and in other situations. Although these methods are sometimes described as ''subjective'' approaches, Block (1961) long ago showed that the combined ratings of observers are largely independent of the idiosyncrasies of any one observer; therefore, when such ratings are aggregated, they are not appropriately characterized as ''subjective.''

Expert Ratings of Breed Prototypes appeared in 18% of the studies reviewed. In these assessments, informants deemed by the researchers to be experts on dogs (e.g., American Kennel Club judges, veterinarians, dog trainers), described, ranked or rated breeds of dogs as a whole rather than specific individual dogs. In these studies, the experts could also make sex-specific judgments. Four of the nine reports included in this review are reanalyses of a single data set (Draper, 1995; Hart and Miller, 1985; Hart and Hart, 1985; Hart, 1995). These data were collected through 96 telephonic interviews, conducted by three veterinary students (Hart et al., 1983; Hart and Miller, 1985). The students asked 48 obedient judges and 48 small-animal veterinarians to compare and rank a selection of seven breeds on 13 questions. When the data were combined, this resulted in the ranking of 56 total breeds on 13 behavioral traits, with 12 independent ratings of each breed on each item.

Observational Tests were used in 16% of the studies. The overall goal of Observational Tests was to assess and describe relatively broad traits discernible in naturalistic environments, thus drawing broader conclusions about the dogs' temperaments and

behavior patterns than is possible using Test Batteries. Like Test Batteries, Observational Tests had two major components: the test itself and the system for scoring the dogs' performance on the test. Unlike Test Batteries, Observational Tests were usually conducted in carefully selected, but not controlled, environments and involved the fortuitous presentation of naturally occurring stimuli. For example, in one study, dogs were walked through a shopping center because it is an uncontrolled public area (Goddard and Beilharz, 1984b). Some Observational Tests also included the presentation of the kinds of experimental stimuli sometimes used in Test Batteries. The target dogs were usually assigned scores on various predetermined temperament traits based on overall observations; for example, in a series of studies, potential guide dogs were judged on cooperativeness based on all behaviors displayed during videotaped walks (Murphy, 1995, 1998).

Some of the studies reviewed (18%) did not fit neatly into any one of these categories because they used combinations of the assessment methods. An example of a study using combined methods was reported by Stephen and Ledger (2003). Dog owners filled out a questionnaire about their dogs' behavior (i.e., Ratings of Individual Dogs) and in a separate phase, unfamiliar testers put the dogs through a series of situations in a controlled environment and rated their behaviors (i.e., Test Battery). In the final step, the researchers compared the scores derived from the two methods.

4.2. Breeds assessed

Another way to summarize the literature is in terms of the breeds assessed. Dogs come in an enormous variety of breeds, with as many as 150 breeds officially recognized by the American Kennel Club (AKC; Registration Statistics, 2004; http://www.akc.org/breeds/reg_stats.cfm) and many others not recognized by the AKC but described elsewhere (Morris, 2002; Wilcox and Walkowicz, 1995). Given this variety, we examined whether the breeds assessed in these dog temperament studies are representative of the breeds that exist or whether there is a bias with some breeds particularly likely to garner research attention. To address this question, the breed composition of the studies is recorded in Table 1.

The Labrador Retriever, Golden Retriever, Beagle, and German Shepherd Dog (GSD) are, respectively, the first, second, third and fourth most commonly registered breeds in the AKC (Registration Statistics, 2004; http://www.akc.org/breeds/reg_stats.cfm). As pure-bred pets and show dogs, they are extremely common. In the studies that reported breed, a staggering 96% of the dogs were purebred. Two of these breeds—the Labrador Retriever and the GSD—were studied particularly frequently. Labradors and GSDs combined dominated the research literature, comprising 32% of the subjects in the studies reviewed. The GSD, which has been surpassed in popularity by the Beagle over the last few years according to the AKC registration records (Registration Statistics, 2004; http://www.akc.org/breeds/reg_stats.cfm), was the most frequently tested breed, comprising 26% of the dogs tested (9208 dogs). Some studies examined huge numbers of these dogs. For example, Reuterwall and Ryman's (1973) study involved 958 GSDs, tested at the Army Dog Training Center of Sollefteå, Sweden. The Labrador Retriever, the most commonly registered breed in the AKC, is the second most frequently tested breed, comprising 6% of the subjects. They too were occasionally present in large numbers in single studies. For

example, 731 Labradors were in Goddard and Beilharz (1982–1983) study of animals with the Royal Guide Dogs for the Blind Association of Australia.

As shown in Table 1, dog temperament assessment studies did not always rely on purebred dogs. Some of the dogs studied were the planned offspring of two purebreds of different breeds. In the studies reviewed, intentional crosses included 16 dogs evenly divided among all possible combinations of Labrador, GSD, Boxer, and Kelpie (Goddard and Beilharz, 1984a, 1986), and 145 Labrador/Golden Retriever crosses (140 in Serpell and Hsu, 2001; 5 in Murphy, 1995).

Also represented in the studies were less common purebred dogs (e.g., Bernese Mountain Dogs; Roll and Unshelm, 1997), and unintentional or unknown mixes of breeds. These studies are different from those not reporting breed in that they make clear that the dogs' involved are not just purebreds of unreported breed, but are actually mixed breeds. Only five studies reporting breed examined unintentional or unknown mixes, totaling 828 dogs (Goodloe and Borchelt, 1998; Hsu and Serpell, 2003; Seksel et al., 1999; Wahlgren and Lester, 2003; Weiss and Greenberg, 1997). Of these, 809 of them were in three studies using Ratings of Individual Dogs (Goodloe and Borchelt, 1998; Hsu and Serpell, 2003; Wahlgren and Lester, 2003). Of the remaining, 10 were in a Test Battery which also included 50 purebred dogs (Seksel et al., 1999), and 9 were in a study composed entirely of mixed breeds (Weiss and Greenberg, 1997).

Are some method-breed combinations more common than others? The breakdown of breeds by assessment method is clearly not random. The most salient patterns appear where huge numbers of dogs are assessed. For example, at least one-third (8792 total dogs) of the dogs in Test Battery studies are GSDs (the most commonly assessed breed overall), tested for their potential as police and working dogs. More than 75% of all dogs in Observational Testing studies are Labrador Retrievers (the second most commonly assessed breed overall), tested for their potential as guide dogs (831 out of 1089 dogs).

4.3. Purpose of study

Not surprisingly, given the diversity of fields doing research on dog temperament and personality, the studies reviewed varied widely in their goals. These goals included determining the suitability of a dog for guide-type work, selecting breeding stock for police dog training centers, and assessing pet dogs' Fearfulness levels.

Ten of the studies reviewed focused on determining the suitability of a dog for guide-dog service work. For example, Goddard and Beilharz (1984a) devised a study to attempt to predict adult Fearfulness in potential guide dogs from tests conducted while they were still puppies.

Three studies aimed to determine suitability for police work and three others focused on suitability for related tasks, such as field work or tracking. For example, a Test Battery was developed for predicting adult police dog effectiveness from the performance of approximately 2-month-old puppies at the South African Police Service Dog Breeding Centre (Slabbert and Odendaal, 1999). This Test Battery included crossing obstacle courses, retrieving objects, novel and startling visual and auditory stimuli, and situations attempting to provoke aggressive behavior. High scores on the retrieval test at 2 months and the aggression test at 9 months significantly predicted success as an adult police dog.

Three of the studies focused on determining the factors involved in aggressive behavior. For example, one study used Ratings of Individual Dogs to investigate whether red and golden ECSs display more aggressive behaviors than do other black and multi-colored ECSs (Podberscek and Serpell, 1996).

The goal of some puppy-temperament assessment methods was to help potential puppy buyers or adopters in selecting a suitable breed and a suitable individual puppy for themselves and their families. There are two types of assessment for this purpose: the breed profile created from Expert Ratings of Breed (e.g., Coren, 1995; Hart and Miller, 1985; Hart and Hart, 1985), and the puppy-behavior test, a type of Test Battery, to be performed by the puppy buyer (e.g., Campbell, 1972; examined in Beaudet et al., 1994).

A handful of other studies have scattered purposes, including developing assessment tools for screening dogs for the presence or prevalence of behavior and temperament problems (Goodloe and Borchelt, 1998; Hsu and Serpell, 2003; Serpell and Hsu, 2001), evaluating previous tests (Beaudet et al., 1994; Weiss and Greenberg, 1997), evaluating the presence of personality traits in dogs (Draper, 1995; Gosling et al., 2003a; Royce, 1955; Svartberg and Forkman, 2002), predicting post-adoption behavior problems in shelter dogs (Hennessy et al., 2001), and determining the relationship between physical build and temperament traits (Keeler, 1947; Lester, 1983).

4.4. Age at testing

As noted above, the goal of many studies has been to predict adult behavior from puppy-temperament. This implies an age-related bias in the studies. To examine the extent of this bias, it is instructive to organize the studies in terms of the age at which the dogs were assessed. To facilitate this goal, in the text and tables we have converted the age information to a common metric of months. Over 20% of assessments in this review were performed for the first time when the dogs were puppies between .23 months (i.e., 1 week) and 6 months of age. Eight were performed when the dog was between 10 and 24 months. Six of those that first assessed the puppies at 6 months old or younger also assessed the dogs on multiple subsequent occasions, with a final test at 12–24 months old; in these studies, researchers tried to use scores from the puppy tests to predict behavior or aptitude when the dog was older (e.g., Wilsson and Sundgren, 1998).

The studies that tested dogs only once tended to test the dogs when they were older. Ten of these studies reported the age at which their first assessment took place as older than 6 months, and of these, eight were 12 months or older. Age is reported in only three of the Ratings of Individual Dogs studies (Hsu and Serpell, 2003; Podberscek and Serpell, 1996; Serpell and Hsu, 2001), and is not discussed in any of the Expert Ratings of Breeds.

Overall, there is a strong tendency towards testing puppies and young dogs. Tests of adult dogs were typically of dogs who were barely adults at just a few years old. A single study did examine dogs with a mean age of 62.2 months (Hsu and Serpell, 2003), and two other studies report ages ranging up to 120 and 204 months (Mahut, 1958; Podberscek and Serpell, 1996, respectively). However, these studies do not affect the overall mean age, which is still less than 24 months. Thus, one striking pattern to emerge is the tendency of researchers to examine young dogs, usually no more than a few years old.

4.5. Breeding and rearing environment

Our review reveals an interesting pattern in terms of the composition of breeding and rearing environments. More than one-third of the studies in our review focused on dogs bred and reared for particular programs. Many of these programs, such as the Swedish Dog Training Center (SDTC), Jackson Laboratories, the Australian Guide Dog Association, and the American Guide Dog Association, attempted to select dogs for breeding. The effects of this temperament-based selective breeding can be seen in various programs. For example, selective breeding based on puppy test performance scores at the Guide Dogs for the Blind training center in San Rafael, California (Scott and Bielfelt, 1976), lead to an improvement in puppy test scores over successive generations; interestingly, this increase in puppy test scores was not matched in the rates at which adult dogs became successful guide dogs, suggesting the puppy tests may not be an ideal criterion for selecting guide dogs, at least in this high-functioning group of subjects.

Many of the dogs in these studies are purebred dogs living as privately owned pets or show dogs. Others were bred to be guide dogs, police dogs, other working dogs, or as research subjects. Only a minority of the dogs studied were from the large populations of rescued and shelter dogs that hope to benefit from temperament research. A disproportionately large number of the dogs examined were dogs specially bred and specially trained for specific working programs. This is an important point to be borne in mind by people seeking to use the research on temperament to understand and predict the behavior of pet or shelter dogs.

4.6. Sexual status of subjects

As noted above, many of the dogs assessed were from programs seeking to breed dogs suitable for specific tasks, such as guide work or police work, and most of the privately owned dogs were intact. Thus, most animals were not spayed or neutered and the effects of castration were addressed in only a few studies. The rare studies that assessed the effects of castration indicated that intact male dogs were the most likely to show aggressive behavior, and intact female dogs were the least likely (Podberscek and Serpell, 1996; Roll and Unshelm, 1997). Podberscek and Serpell's study also revealed that neutering an adult dog in reaction to his aggressive behavior does not reduce future aggression. Overall, however, we know little about the effects of spaying and neutering on dog temperament in general, and even less about how the animal's age at castration affects its later temperament. With the increasing prevalence of laws requiring spay and neuter surgeries before a pet dog can be adopted from a shelter or rescue and the prevalence of spayed and neutered dogs in our daily lives, the effects of these surgeries on temperament is another area needing more research.

4.7. Summary of general survey

To provide some coherence to the enormously varied work on dog temperament, we organized the literature in terms of six frameworks. Organizing the studies in this way allowed us to make several observations about the state of the field. First, there is great

diversity in most components of the research, including such features as the goals and the disciplinary bases of the studies. Second, the studies can be usefully categorized in terms of four assessment methods (Test Batteries, Ratings of Individual Dogs, Expert Ratings of Breed Prototypes, and Observational Tests). Third, most of the dogs studied (90%) were purebred, with Labrador Retrievers and GSDs composing 32% of the subjects. Only five studies reported examining unintentional or unknown mixes, totaling only 828 dogs (Goodloe and Borchelt, 1998; Hsu and Serpell, 2003; Seksel et al., 1999; Weiss and Greenberg, 1997; Wahlgren and Lester, 2003). Fourth, there is a systematic pattern in which certain breeds are associated with particular types of studies; more than one-third (8792 total dogs) of the dogs in Test Battery studies were GSDs, and more than 75% of all dogs in Observational Testing studies were Labrador Retrievers. Consequently, very few breeds other than Labrador Retrievers have been examined by Observational Testing. Fifth, there is a tendency in the research towards testing puppies and young dogs, with older adult dogs (over 4 years old) infrequently studied and elderly dogs almost entirely neglected by the research literature. Sixth, most of the studies in our review focused on dogs bred and reared for particular programs while tests selecting dogs as pets (e.g., from shelters or rescues) were rare. And last, most dogs involved in these studies were not spayed or neutered and the effects of castration were addressed in only a couple of studies.

5. Review and evaluation of the empirical findings

Our review has identified enormous variability in the field in terms of the types of assessments, research purposes, and other attributes of the studies themselves. We next extend our review to the substantive findings of the studies. Specifically, we ask what traits have been studied and we evaluate the evidence for the reliability and validity of the assessment methods developed so far.

5.1. What temperament traits have been studied in dogs?

To determine which traits have been identified in studies of non-human animals, Gosling and John (1999) reviewed the structural studies of temperament and personality in non-human species, ranging from chimpanzees to octopuses. They found evidence for several basic dimensions that recurred across species, with especially strong cross-species evidence for Anxiety/Nervousness, Sociability, and Aggression. What can we learn from the present, more focused review of the temperament traits that have been studied in dogs? In this section, we describe the findings of a systematic analysis of the traits and behaviors examined in past dog research.

Dog temperament researchers have studied a broad array of traits ranging from gun shyness to snapping at children. Summarizing these findings is not a straightforward task because, as we saw above, the studies used different methods, different populations, and are grounded in different disciplines, resulting in a non-standardized vocabulary. On occasion, the same term was used to refer to different behaviors. For example, in one study ‘‘temperament’’ was defined as ‘‘character, sensitivity, discrimination, spirit, and intellect’’ (Slabbert and Odendaal, 1999), in another study as ‘‘a combination of underlying traits’’

(Humphrey, 1934), and in yet another study as ''physical flexibility and intensity of reaction to different environmental stimuli'' (Ruefenacht et al., 2002). In addition, different terms have been used to refer to very similar behaviors. For example, in one study a dog that ''goes up to any stranger on sight and makes friends'' was scored as high on ''Confidence'' (Humphrey, 1934, p. 133), but the similar behavior of exhibiting ''friendly greetings to strangers'' (with friendly tail-wagging, for example), was scored by other researchers as high on ''Friendliness'' (Goodloe and Borchelt, 1998) or ''Sociability'' (Hennessy et al., 2001). In short, no standard lexicon of dog traits and behaviors exists, with the result that traits and behaviors have not been defined consistently across studies. The idiosyncratic terms used in the different studies impede attempts to make cross-study comparisons of what has been learned.

There is clearly a need to develop a common language with which it describes canine temperament. Despite an attempt by Goodloe and Borchelt (1998) to develop a standard lexicon of dog traits and behaviors, none have yet been widely adopted. Therefore, to allow us to summarize the findings across all articles, we used a systematic procedure in which expert judges categorized the varied constructs with a standardized set of terms. Our procedure involved three major steps.

5.2. Step 1: extracting behavioral descriptions

The first step was to gather the descriptions of the behavior that had been studied but to avoid any biases introduced by the researchers' idiosyncratic choice of labels. In each study, we located the descriptions of the behavior and wrote the descriptions on index cards with one index card for each behavior. The descriptions of behaviors provided in the articles varied in the detail of the descriptions and the degree to which the descriptions included trait-related terms. We removed terms indicative of the dogs' internal motivations or emotional states and terms directly referring to traits, such as ''fearful,'' ''timid,'' and ''curious.'' These left behavioral descriptions that were less biased by the researchers' theories about which traits underlie the behaviors. For example, instead of ''Social attraction,'' a term used by Campbell (1972) and later by Beaudet et al. (1994), our card would be based purely on the behavioral descriptions provided by the researcher: ''a puppy's tendency to move towards a human tester who has placed the puppy in a corner of an observation area, moved to the opposite corner, crouched, and clapped his/her hands quietly.''

It was sometimes impossible to separate descriptions of behavior from labels describing a temperament trait. For example, although Mahut (1958) reports making detailed, objective notes about dogs' behavior, all that is reported is the classification of these notes into categories such as ''curiosity'' and ''wariness.'' We were unable to extract more detailed descriptions, so we used these non-descriptive classifications in our index card task.

This procedure resulted in a total of 623 different index cards. The index cards were assigned code numbers associated with the article from which they were drawn but the key to the code was not known by the judges. This ensured that the judges in Steps 2 and 3 were blind to the identity of the researchers and study from which the descriptions were taken.

5.3. Step 2: development of temperament categories

The first author and a research assistant/veterinary technician served as judges in a sorting task designed to identify the major temperament dimensions. Both judges had a moderate amount of professional work and research experience (at least 5 years each) with dogs. The cards were shuffled and the judges were instructed to sort them into groups based on the temperament traits likely to be underlying the behaviors described. For example, the cards displaying ''Is 'spooked' by odd or unexpected things or objects'' (from Serpell and Hsu, 2001) and ''avoids or is fearful of unfamiliar children'' (from Goodloe and Borchelt, 1998) were placed, by both judges, together in a single pile. The judges were under no time pressure. Judges were allowed to place one behavior in more than one pile to indicate that the behavior is potentially related to more than one underlying temperament dimension. To do this, the judges copied the code number from the back of the relevant index card onto a new index card and placed cards in each pile; the same process was repeated if a description was deemed to fit into more than two categories.

Once all the cards had been grouped in this way, there were seven piles, with 92% agreement across the two judges. The two judges worked together to provide consensual labels for the seven piles. The final labels were Reactivity/Excitability–Stability, Fearfulness–Courage/Confidence, Aggression–Agreeableness, Sociability/Friendliness–Lack of Interest in Others, Openness–Non-Openness (later renamed Responsiveness to Training), Dominance–Submission, and Activity level.

5.4. Step 3: classification of behaviors by a panel of expert judges

To ensure these categorizations were not attributable to the original judges' idiosyncratic experiences, we designed a second categorizing task undertaken by seven additional expert judges. The panel of judges were selected on the basis of their experience with dogs, the variation of situations in which they observed dogs, and the number of years they had worked with dogs. The final panel was composed of a veterinarian, a public-shelter dog temperament tester, three dog trainers with varying specialties, a professor studying animal social behavior, and a graduate student studying dog behavior. They had between 7 and 20 years of experience working with dogs and at least 3 years of formal education in canine or animal behavior. Only the temperament tester specialized in researching or assessing temperament.

All of the expert judges were given identical sets of 623 index cards and sorting instructions. They were also given the list of the seven temperament dimensions derived in the previous step. To allow the judges to disagree with the classifications provided by the judges in Step 2, there were two additional categories labeled as Other and Not Temperament-Related. The expert judges were told to take their time in separating the cards into groups corresponding to the nine categories. The judges were told that a behavioral description written on any one index card could be indicative of more than one temperament dimension or ambiguous as to the underlying dimension leading to the behavior. In such cases, the experts were instructed to copy the number from the back of the relevant index card onto a new card and place the trait in two temperament dimension piles; the same process could be used if a description fit into more than two temperament

dimension categories. If the judges thought the behavioral descriptions did not fit into any of the seven temperament dimensions, the judges were instructed to place the card in the category ''Not Temperament-Related'' or ''Other'' and provide an explanation for why they had selected this category.

The results of this Expert Temperament Categorizing task were reassuringly consistent across the expert judges. Average pair-wise agreement across judges was 89%, with a maximum agreement between two judges of 95% and a minimum agreement of 80%. Points of disagreement among judges included what dimensions underlie the traits Barking, Problem Solving, and Fearfulness. Typically, the more detail present on the index cards, the more agreement among judges. For example, judges were in less agreement about how to categorize ''Barking'' than on how to categorize ''Barks and sometimes growls when approached by a male stranger.''

As noted above, the seven judges in Step 3 were at liberty to disagree with the categorizations developed by the two judges in Step 2. An inspection of the Step 3 judges' categorizations showed that they did indeed disagree with a distinction made in Step 2. In particular, the panel of seven dog experts saw less distinction between the Reactivity/Excitability dimension and the Fearfulness dimension than between the other dimensions, at least in the context of temperament-testing studies. Cards were quite frequently categorized as falling into both the Reactivity and Fearfulness categories. This overlap of dimensions is consistent with research in the human domain, where Reactivity and Fearfulness are components of the same Emotional Stability dimension. Further investigation of the dimensions of Reactivity and Fearfulness in dogs would need to be conducted for us to know whether the two are indeed independent, or whether they might fall under an even broader super-ordinate category.

5.5. *Potential limitations of sorting method*

Although our multi-stage procedures were designed to reduce the impact of any single judge and are consistent with standard meta-analytic procedures (Lipsey and Wilson, 1996; Rosenthal, 1991), and very similar procedures have been utilized in various other meta-analyses related to personality in humans (e.g., Barrick and Mount, 1991; Bogg and Roberts, 2004; Heller et al., 2004), it is important to recognize the limitations of this method. One potential limitation is the possibility that the results are influenced by idiosyncratic experiences of the judges, such that a different group of judges might produce different results. Another potential limitation is that the labels generated in Step 2 could have biased the sorting task in Step 3; specifically, the choice of labels in Step 2 could have influenced the views or limited the options of the judges in Step 3. To safeguard against these potential limitations, we implemented multiple safeguards. First, we made it very clear to the seven judges in Step 3 that the category labels with which they were provided were merely suggestions, so the Step 3 judges could choose not to use these labels if the labels were inadequate or inappropriate. We also provided the judges with ''None'' (or ''Not Temperament-Related'') and an ''Other'' category for cards that did not fit into the categories suggested in Step 2. After the judges had completed the sorting task, we asked them to describe each of the temperament dimension categories so that we could be certain that the judges were using the labels similarly. Reassuringly, 59 of the 63 descriptions given

by these Step 3 judges were almost exact matches to those Step 2 judges had used when they selected labels for the categories.

Of course, there is a danger that our safeguards would not be effective if the Step 3 judges felt they could neither use the ''Not Temperament-Related'' and ''Other'' categories nor generate their own categories. However, the results of the sorting task showed the judges were willing to use these two categories. An analysis of the frequencies with which the Step 3 judges used the various categories in the sorting task showed they used these two categories almost as frequently as they had used the other seven categories.

This frequency of use suggested both that the judges in Step 3 were comfortable using the categories, and that they agreed with the judges in Step 2 that some of the traits studied were simply not temperament traits (e.g., body sensitivity). In addition, as noted above, four of the seven judges questioned the Step 2 judges' separation of Reactivity and Fearfulness, suggesting the Step 3 judges were not constrained by the categories generated by the Step 2 judges. These four judges recommended the two categories be combined and relabeled as ''Nervousness'' or ''Nerve Stability.'' This recommendation demonstrates that the judges took the provided labels as suggestions and not as final labels.

In addition, we implemented the following four safeguards against the danger of generating idiosyncratic categories. First, when selecting judges for Step 2, we strove to identify judges with different kinds of professional experience with dogs. Second, we had these two judges complete their sorting task entirely independently and, if discrepancies arose, discuss them until consensus was reached. This limited the impact of each judge on the results and safeguarded against the categories the judges generated being unique to this analysis. Third, when selecting the group of seven judges to participate in Step 3, we again strove to build a group with diverse professional experience with dogs. Fourth, this group of judges also completed the sorting task entirely independently from one another. In addition, both sets of judges were under no time constraints. Despite these safeguards, we acknowledge it is still possible that our results might be unique to this group of judges and we caution the reader to interpret the findings with these caveats in mind.

5.6. Results from the sorting task

Table 2 summarizes the results of our analyses. For this summary, we have combined the ''Not Temperament-Related'' and ''Other'' categories. Thus, the eight column headings show the eight categories identified in our analyses. We relabeled the Openness category as Responsiveness to Training to avoid confusion with Openness as defined in the human-personality literature. As shown in the table, Reactivity, Fearfulness, Sociability, Responsiveness to Training, and Aggression have been examined more frequently than the other dimensions.

Traits related to the Reactivity dimension were studied quite frequently, in 39 of the studies in our review. High Reactivity was indexed by such behaviors as repeated approach/avoidance of novel objects, raised hackles, and increased activity in novel situations. Low Reactivity was characterized by such behaviors as a relative lack of change of behavior in new situations, and approaching novel stimuli without backing away. In the tests, Reactivity was assessed through such procedures as presenting a novel object or series of novel objects to a puppy and recording its subsequent behavior (Goddard and Beilharz,

Table 2
Which traits have been studied in dogs?—a review of past research

	Traits							
	Reactivity	Fearfulness	Activity	Sociability	Responsiveness to Training	Submissiveness	Aggression	None/Other
Beaudet et al. (1994)			Activity level	*Following* Social attraction	*Following*	*Restraint dominance* *Elevation dominance* *Social dominance*		*Restraint dominance* *Elevation dominance* *Social dominance*
Bradshaw and Goodwin (1998)	Reactivity		*Immaturity*	*Immaturity*	Housetrainability		Aggressivity	
Campbell (1972)				*Following* Social attraction	*Following*	*Restraint dominance* *Elevation dominance* *Social dominance*		*Restraint dominance* *Elevation dominance* *Social dominance*
Cattell et al. (1973)	Calmness Emotionality Excitation	Timidity Apprehension	Exuberance	Self-sufficiency Aloofness	Obedient cooperation		Aggressiveness	Competence
Cattell and Korth (1973) [a]	EII (Social reactivity) *EIII* (*Affective arousal*)	EI (Extraversion) EVIII (Apprehension) *EIII* (*affective arousal*)		AII (Cooperation) *EIV* (*Independence*)	AI (Un-named) EVI (Calmness)	EV (Timidity) *EIV* (*Independence*)		AVI (Breed aptitude) EVII (Un-named)
Coren (1995)	Sound reaction Stability Reaction to novel stimuli			Social Attraction (Approaching, Following)	Willingness to work (Retrieval)	Social dominance (Restraint, Forgiveness, Loss of control)		Touch sensitivity response to Food incentive
Draper (1995)	*Reactivity–surgency*			*Reactivity-surgency*	Trainability-Openness	*Aggression–Non-agreeableness* [*Dominance over owner*]	*Aggression–Non-agreeableness* [*Dominance over owner*]	Investigation
Goddard and Beilharz (1986)	*Fearfulness*	*Fearfulness* (Approach/avoid)						
Goddard and Beilharz (1985)		Confidence *Aggression–Dominance* (*Hackles, Biting*)				Submissiveness *Aggression–Dominance*	*Aggression–Dominance*	
Goddard and Beilharz (1984a)	*Fearfulness*	*Fearfulness*						

Goddard and Beilharz (1984b)	PC3[b] (Fearful and *Excitable*)	PC1 (General fearfulness)	PC2 (Activity on walk) PC5 (Activity in home, on free run) PC6 (Activity in home)		PC4 (recall) *PC7* (*Repetitions of name during recall*)			*PC7* (*Repetitions of name during recall*)
Goddard and Beilharz (1982–1983)	Distraction Sensitivity *Fearfulness and High activity*	Fearfulness *Fearfulness and High activity*					*Nervous aggression* Aggression	General performance sensitivity (body, sound)
Goodloe and Borchelt (1998)	*Barking 1* *Barking 2*	*Fear/Avoidance of strangers*		*Fear/Avoidance of strangers* Friendliness Affiliation	Play 1 *Compliance*	Submission *Compliance* *Mounting other dogs*	Aggression to family/strangers/ unfamiliar dogs Biting	*Barking 1* Separation vocalization Play 2 Play 3 Destruction Digging/burying Eating sensitivity Male-related behaviors *Mounting other dogs* Mounting objects
Gosling and Bonnenburg (1998)	Disorganized/ Organized Relaxed Temperamental Touchy Moody [*Rude*]	Withdrawn Fretful Nervous Anxious	Quiet	Withdrawn Cold/Warm Extraverted Unkind/Kind Shy Untalkative/Talkative Verbal Bashful	Considerate Cooperative Prompt	Bold Bashful	Kind/Unkind	Artistic Careless Complex Uncreative/Creative Deep Inefficient/Efficient Harsh Imaginative Intellectual Unintelligent Unenvious/Jealous Philosophical Practical [*Rude*] Sloppy Unsympathetic/ Sympathetic

Table 2 (*Continued*)

	Traits							
	Reactivity	Fearfulness	Activity	Sociability	Responsiveness to Training	Submissiveness	Aggression	None/Other
Gosling et al. (2003a)	Neuroticism			Extraversion	Openness		Agreeableness	
Hart et al. (1983); Hart and Miller (1985); Hart (1995)	Excitability	*Snapping at children*	General activity	*Snapping at children*	Obedience training	[*Dominance over owner*]	[*Dominance over owner*]	Destructiveness
	Excessive barking			Affection demand	Playfulness Housebreaking ease		Snapping at children Territorial defense Aggressive to dogs	Watchdog barking
Hart and Hart (1985)	*Reactivity*	*Reactivity*		*Reactivity* (Affection demand)	Trainability Playfulness	[*Dominance over owner*]	Aggression [*Dominance over owner*]	Destructiveness
Hennessy et al. (2001)		Flight Wariness *Timidity*	Locomotor activity	Sociability *Timidity*				Solicitation
Hsu and Serpell (2003)	Excitability	*Stranger-directed fear* *Dog-directed fear or aggression* Non-social fear		Attachment or Attention-seeking behavior	Trainability		*Stranger-directed aggression* *Dog-directed fear or aggression* *Chasing* Owner-directed aggression	Separation-related behavior *Chasing* Pain sensitivity
Humphrey (1934)	*Energy*	*Confidence*	*Energy*	*Confidence* (approaching to make friends)	Nose ability *Intelligence* *Willingness*	Self-right	Sharpness Fighting (own kind)	Sensitivity (body, ear) *Intelligence* *Willingness*
James (1951)						Dominance over other pups Guarding food area		Which pups' company each prefers
Keeler (1947)		Nervous Courageous						Agile Tame
Ledger (2003)							Aggression	

Ledger and Baxter (1996, 1997)	Excitability	Timidity			Obedience	Aggression	Separation-related problems
Lester (1983)	Lethargic [*Emotional*]	Timid	Lethargic Active	Friendly	Easy to train Obedient	Aggressive	Curious [*Emotional*]
Mahut (1958)	[*Fearfulness*]	[*Fearfulness*]			[*Fearfulness*] (coming if called by mask-wearer) Interest in stimuli		
Murphy (1998, 1995)	*Low concentration* *Dog distraction* Excitability	Anxiety Suspicion Nervousness			*Low concentration* *Dog distraction* Low willingness	Pure aggression Nervous aggression Dog aggression	Low body sensitivity Immaturity
Netto and Planta (1997)		*Aggression*				*Aggression*	
Podberscek and Serpell (1996)						Aggression	
Reuterwall and Ryman (1973)	Adaptiveness to different situations and environment *Ability to meet with sudden, strong auditory disturbances*	*Ability to meet with sudden, strong auditory disturbances*		Affability *Disposition for fighting in a playful manner*		Disposition for self-defense Disposition for handler defense	*Disposition for fighting in a playful manner* *Disposition for forgetting unpleasant incidents*
Roll and Unshelm (1997)							Aggression
Royce (1955)	Heart reactivity Autogenic reactivity *Activity level*	Timidity I Timidity II (with Withdrawl)	*Activity level*			Aggressiveness	4 un-labeled traits
Ruefenacht et al. (2002)	[*Reaction to gunfire*] Nerve stability Hardness	[*Reaction to gunfire*] Self-confidence			Temperament	Sharpness Defense drive *Fighting drive*	*Fighting drive* (tolerating stick beats)
Serpell (1983)	Excitability Reaction to owner's homecoming	Nervousness		Friendliness to strangers Friendliness to other dogs Loyalty/Affection Sensitivity to owner's moods Expressiveness	Obedience on walks Obedience at home Attentiveness	Territorial barking Protectiveness Possessiveness	Attitude on walks Attitude about food Intelligence/Aptitude Reaction to separation Sense of humor

Table 2 (*Continued*)

	Traits							
	Reactivity	Fearfulness	Activity	Sociability	Responsiveness to Training	Submissiveness	Aggression	None/Other
				Attachment (*1 person*)				*Attachment* (*1 person*)
Serpell and Hsu (2001)	[*Chasing*]	*Dog-directed fear/aggression* Non-social fear *Stranger-directed fear/aggression*	Energy level	*Stranger-directed fear/aggression* *Attachment* (*1 person*)	Trainability		*Stranger-directed fear/aggression* Owner-directed aggression *Dog-directed fear/aggression* [*Chasing*]	*Attachment* (*1 person*)
Slabbert and Odendaal (1999)	*Startle test* *Gunshot test*	*Startle test* *Gunshot test*		[*Retrieval test*]	[*Retrieval test*] Obstacle test		Aggression test	[*Retrieval test*]
Stephen and Ledger (2003), also Stephen et al. (2001)	Excitability	Timidity		Playfulness	Obedience		Aggression	
Svartberg (2002)	*Boldness/Shyness*	*Boldness/Shyness*		*Boldness/Shyness*				*Boldness/Shyness*
Svartberg and Forkman (2002), also Saetre et al. (in press)	*Curiosity/fearlessness*	*Curiosity/fearlessness*		Sociability				
	Chase-proneness			*Playfulness*			Aggressiveness *Chase-proneness*	*Playfulness*
van der Borg et al. (1991)		Fear Fear-induced aggression			*Disobedience*		Aggression	*Disobedience* (due to lack of training) Separation anxiety Miscellaneous behavior
Wahlgren and Lester (2003)		Factor IV (Timid, Emotional)		*Factor I* (*high Sociability and Friendliness; low Aggression and Bad-temperedness*)	Factor II (Curious, Active, Independent) Factor III (Obedient, Clever, Protective)		*Factor I* (*high Sociability and Friendliness; low Aggression and Bad-temperedness*)	
Weiss and Greenberg (1997)	*Attention/Distraction* Excitement	*Fear/Submission*			*Attention/Distraction*	*Fear/Submission* Dominance		

Wilsson and Sundgren (1997)	Nerve Stability Hardness *Prey drive*	Courage	[*Temperament*]	Affability	Cooperativeness [*Temperament*]	Sharpness Defense drive *Prey drive*
Wilsson and Sundgren (1998)	Nerve Stability Hardness *Prey drive*	Courage	[*Temperament*] Energy level	Affability	Cooperativeness [*Temperament*]	Sharpness Defense drive *Prey drive*

Note. All dimension labels are those used by the authors. The study authors' definitions of temperament have been used, so we have not excluded items that would not normally be considered temperament constructs (i.e., specific behaviors). Those traits that fell into more than one category are italicised. We list in square brackets those traits that did not elicit 100% agreement among the expert judges in terms of category membership. We provide in standard brackets, where appropriate, more information about traits.

[a] The authors listed and described, but did not always label, the factors derived from their analyses.

[b] PC indicates Principle Component.

1986). The labels and descriptors given to this dimension by researchers included ''excitability'' (Goddard and Beilharz, 1982–1983; Hart et al., 1983; Hart and Miller, 1985; Hart, 1995; Hsu and Serpell, 2003; Ledger and Baxter, 1996, 1997; Murphy, 1995; Serpell, 1983; Stephen et al., 2001; Stephen and Ledger, 2003), ''sound reaction'' (Coren, 1995), and ''heart reactivity'' (Royce, 1955).

Fearfulness was also studied frequently, in 43 studies, and frequently overlapped with Reactivity. One possible reason for this is that dogs may exhibit similar or indistinguishable behaviors as a result of differing emotional states. A dog may exhibit signs of excitement, pacing or running around, approaching objects and then avoiding them, Barking, and so on making it difficult to decipher behavioral reactions due to Fearfulness versus Reactivity (Hoffman, 1999). According to our sorting task, shaking and a tendency to avoid novel stimuli without approaching them are associated with high levels of Fearfulness. In the temperament tests, Fearfulness was often assessed by recording reactions to novel stimuli or situations (Murphy, 1995, 1998). Low levels of Fearfulness were sometimes labeled as ''Courage'' (Wilsson and Sundgren, 1997, 1998; Reuterwall and Ryman, 1973), ''Confidence'' (Goddard and Beilharz, 1985; Humphrey, 1934), and ''Self-confidence'' (Ruefenacht et al., 2002). Some labels given to Fearfulness include ''Apprehension'' (Cattell and Korth, 1973), ''Dog-directed Fear or Aggression'' (Hsu and Serpell, 2003; Serpell and Hsu, 2001), and ''Timidity'' (Hennessy et al., 2001; Ledger and Baxter, 1996, 1997; Royce, 1955; Stephen et al., 2001; Stephen and Ledger, 2003; Wahlgren and Lester, 2003).

Sociability was studied quite frequently, in 31 studies. The traits categorized under this dimension were also sometimes categorized under Responsiveness to Training, perhaps because interest in people is central to Sociability and to interest in training. Sociability was indexed by such behaviors as initiating friendly interactions with people and other dogs. In temperament tests, Sociability was primarily assessed by setting up a meeting between the dog and an unfamiliar person (Humphrey, 1934) or dog (Goddard and Beilharz, 1986). Sociability was given a variety of different labels by researchers, including ''Extraversion'' (Gosling et al., 2003a), ''Affection demand'' (Hart, 1995), and ''Affability'' (Reuterwall and Ryman, 1973).

Responsiveness to Training was studied in 34 of the articles reviewed, and was indexed by such behaviors as working with people, learning quickly in new situations, playfulness, and overall reaction to the environment. This dimension seems very closely related to the dog's tendency to stay focused and engaged in a given activity, and thus, may be more suitably termed ''Distractibility'' or ''Focus.'' The trait was assessed through such procedures as giving puppies puzzles to solve (Cattell and Korth, 1973) and ''willingness'' to work with a person (e.g., Goddard and Beilharz, 1982–1983). Tests for this trait vary from specific to broad. For example, a very specific test was the retrieval test, said to be a test of how willing a puppy is to comply with a human in going to get an object and then returning with it (Slabbert and Odendaal, 1999). In contrast, a broad method of assessing Responsiveness to Training, labeled ''temperament,'' examined the dogs' reactions over a variety of situations, looking for whether the dog exhibited varying reactions and interest in its environment (Ruefenacht et al., 2002). Low Responsiveness to Training was a lack of cooperation or responsiveness to training, or a lack of interest in the situation, while high Responsiveness to Training was the reverse. Labels used to define Responsiveness to

Training included ''problem solving'' (Cattell and Korth, 1973), ''willingness to work'' (Coren, 1995), and ''cooperative'' (Gosling and Bonnenburg, 1998).

Aggression was studied in 30 of the articles reviewed. It was indexed by behaviors such as biting, growling, and snapping at people or other dogs. These behaviors could also be caused by fear and may be performed in self-defense. In such cases, the trait is also related to Fearfulness, but reflects a specific aggressive response to a fearful stimulus. The more dramatic testing procedures used to assess Aggression included such activities as having strangers approach and attack either the dog or the dog's handler (Reuterwall and Ryman, 1973; Ruefenacht et al., 2002). Aggressive behavior was also sometimes divided into subcategories, or types of Aggression, usually on the basis of the cause of the Aggression. For example, Aggression was divided into categories representing Aggression in the service of dominance (Goddard and Beilharz, 1985) and Aggression as a result of Nervousness (Goddard and Beilharz, 1982–1983). Other studies divided types of Aggression by targets, such as stranger directed fear/aggression, owner-directed aggression, dog-directed fear/aggression (Serpell and Hsu, 2001). Also, in studies looking for dogs that can work as police dogs a very specific subset of Aggression is tested; it was called ''sharpness,'' and defined as the willingness to bite a human being (Humphrey, 1934; Ruefenacht et al., 2002; Wilsson and Sundgren, 1997, 1998).

There is some debate about whether the Dominance and its opposite, Submission, should be considered a trait or a social outcome (Gosling and John, 1999). Nonetheless, behaviors reflecting this dimension were present in 16 of the articles reviewed. Dominance was reflected in such behaviors as refusing to move out of a person's path, or ''self-right'' (Humphrey, 1934). Dominance and Submission with other dogs was assessed in James' (1951) study of the development of hierarchy in puppies, in which Dominance was judged by observing which dogs bullied others, and which guarded the food area and ate first, and Submission was judged by puppies getting out of a bully's way. Submission was also reflected by such behaviors as urination upon greeting people (Wilsson and Sundgren, 1998). However, clear behavioral definitions are not provided for all conceptualizations of Dominance; we were unable to find clear and specific descriptions of the behaviors meant to characterize a type of Dominance called ''Dominance over owner'' (e.g., Draper, 1995).

Activity was discussed in 15 studies. Activity has often been assessed by placing a puppy or dog in an empty arena with gridlines on the floor and seeing how many times the puppy or dog crosses the lines (see Wilsson and Sundgren's arena test, 1998). The studies offered various labels including ''Activity'' (Cattell and Korth, 1973; Goddard and Beilharz, 1984b; Reuterwall and Ryman, 1973), ''General activity'' (Hart and Miller, 1985), or ''Locomotor activity'' (Hennessy et al., 2001). Activity level significantly drops between 6 and 12 months of age (Serpell and Hsu, 2001). Goddard and Beilharz (1984b) found that puppy general activity level is a weak predictor of adult activity level due to a decrease in activity as age increases. They also found that activity level ''is of relatively little importance compared to traits which control activity in specific situations'' (Goddard and Beilharz, 1984b, p. 275). However, Activity has been identified as a potentially useful, though weak, predictor of adult Dominance/Submission when paired with another test of puppy behavior (Beaudet et al., 1994). This

is because Activity seems to moderate the predictive value of the other traits. Therefore, even if activity level does not directly predict adult outcomes, it may still be useful to assess activity as a potential moderator variable.

The categories of ''Other'' and ''Not Temperament-Related'' are represented in Table 2 as the final column, ''None/Other.'' This category was used for variables examined by 23 different articles, and we have listed each individual variable. The two groups were condensed to one because experts did not identify an additional temperament dimension, but rather classified the behavior as being associated with factors that are not based on temperament. For example, ''disobedience'' (van der Borg et al., 1991) initially appeared as if it fit under Responsiveness to Training, however, in this example, the disobedience was due to a lack of training; the shelter dogs assessed in this study may not have ever been trained to know the commands they were asked to perform during testing so we cannot attribute their lack of obedience to their temperaments. Sometimes, the authors of an article labeled a behavior variable in a way that made it appear to be temperament-related, such as ''dominance tests'' (Beaudet et al., 1994), but our judges (who were blind to the label provided) agreed that the tests were not actually assessing Dominance/Submission. Other examples of variables in this category include ''touch sensitivity'' (Coren, 1995) and ''hearing sensitivity'' (Goddard and Beilharz, 1982–1983).

The fact that the enormous number of terms in Table 2 can be classified into seven categories of temperament underscores the need for a standard language for describing temperament traits and dimensions in dogs. When each author creates a new set of words with which to discuss the same temperament traits, it not only makes comparisons across studies difficult, but is also a process of recreating the wheel. We propose that the seven categories derived from our review of the literature represents a sensible starting point for the development of such a lexicon of canine temperament descriptors.

6. Are assessments of dog temperament reliable?

If temperament tests are to be of any value, they must be shown to be both reliable and valid. Reliability is a prerequisite for validity, and so we review the evidence for reliability first.

The first thing to conclude about reliability is that with the few exceptions we will discuss in more detail, researchers have rarely reported reliability of any kind. Those studies that have examined reliability have done so from a variety of perspectives, using different statistical indices, assessing different types of reliability, and computing these reliabilities differently. We culled from the studies in our review all measures of reliability. Unfortunately, most studies that addressed reliability did so without references to numerical indices. For example, Lester (1983) described inter-judge reliability as ''acceptable'' on all but three traits assessed, but did not specify the standard by which ''acceptable'' was evaluated. Slabbert and Odendaal (1999) discussed reliability in the context of attempts to improve reliability by using trainers (versus dog owners) as raters but they did not provide any numerical indices of reliability. Weiss and Greenberg (1997) had raters train together, prior to temperament testing, in order to ensure what they termed ''high'' inter-rater reliability but again, no numerical reliability standard was reported.

None of these studies could be included in our quantitative review because they did not report reliability numerically.

In addition, we had to exclude from our analyses studies that reported incomplete, incomparable, or unique measures of reliability. For example, although Murphy's (1995) study aimed to explicitly address the consistency with which guide dogs' temperaments could be rated by trained judges, the article did not provide a quantitative index of reliability and we could not compute reliability because the data set provided was incomplete. The data provided were a small subset of the whole sample. Although these data were described as representative, they consisted of only a handful of ratings so we considered them incomplete and did not compute a reliability coefficient.

We also had to exclude some studies that reported correlations between tests and retests in a way that did not address the tests' reliability. For example, Goddard and Beilharz (1986) reported some, but not all, correlations between scores at the various ages at which they assessed guide-dog puppies, making the point that the correlations increased as the dogs aged. The purpose of Beilharz's studies (1986, 1984a,b) was to evaluate this change and to determine how old a puppy must be for the puppy's Fearfulness level to predict its adult Fearfulness. Because these tests sought to index change in the subjects, not repeatability of the test, the scores were not appropriate for our analyses. In addition, reliability could not be computed in the studies that simply obtained frequency estimates of certain behaviors (e.g., Podberscek and Serpell, 1996) or obtained ratings from a single judge (e.g., Goodloe and Borchelt, 1998).

The remaining reliability coefficients that we were able to uncover or compute are reported in Tables 3 and 4. The tables are divided by type of analysis; Table 3 shows the inter-observer agreement and test–retest reliability correlations, and Table 4 shows internal consistency as indexed by Cronbach's alpha.

Table 3 is divided into two types of reliability: inter-observer agreement and test–retest reliability. The studies using inter-observer agreement used the traditional method of analysis in which each variable is analyzed across subjects (instead of computing reliability within subjects). The correlations suggest that inter-judge agreement varies greatly across studies and traits. Although strong agreement is possible, it is by no means guaranteed; the sample-weighted mean agreement correlation was .60, but the agreement correlations ranged from .00 to .86. A study by Goodloe and Borchelt (1998) was excluded from our table because the data are not complete or precise enough to allow us to integrate them into our quantitative analyses. However, their data also support the idea that dog temperament can be assessed reliably. Ninety-six percent of their inter-observer correlations were above .60, and of those 55% were above .80; their lowest correlations were on four items, reported only as less than .50. Goodloe and Borchelt (1998) emphasized the point that dogs may interact with raters differently, which would diminish the apparent reliability, as the dogs may behave consistently with each person, but differently across people. Clearly, given that reliability is a fundamental standard of all measurement studies, future research is badly needed on this possibility and others. In general, whenever appropriate data are available, reliability indices should be reported, as is standard practice in research on human personality.

Two studies appear in the test–retest reliability category, listed in the lower section of Table 3, examining the correlation between scores when dogs were tested twice. One of

Table 3
How reliable are personality measures of dogs?: inter-observer agreement and test–retest reliability

Study	Assessment method[a]	Inter-observer agreement						Sample size	Number of indicators	Retest interval
		Mean correlation	S.E.$_{z_r}$[c]	Maximum		Minimum				
				Correlation	Item label	Correlation	Item label			
Inter-observer agreement computed for each variable across subjects										
Gosling et al. (2003a)	Combination	.62	.12	.76	Extraversion	.55 .55	Agreeableness Openness	78	4[d]	
Goddard and Beilharz (1982–1983)	Observational	.47	.41	.70	Nervousness	.00	Willingness	9	14	
Stephen et al. (2001) (also Stephen and Ledger, 2003)[f]	Combination	.75	.30	.86	Aggressiveness	.49	Excitability	13–14	75[g]	
Unweighted mean		.56		.78		.37				
Sample-weighted mean		.60		.77		.50[e]				
Test–retest reliability										
Goddard and Beilharz (1986)	Combination	.39	.14	.52	Activity 7 and 9 weeks	.21	Activity 5 and 9 weeks	102	1	Variable
Netto and Planta (1997)	Test Battery	.77	.17	.79	Unfamiliar female dominant dog in area[b]	Many non-significant (effect sizes not reported)		37	43	6 months
Unweighted mean		.58		.82		.21[h]				
Sample-weighted mean		.45		.61		.21[h]				

[a] The categories used here refer to the types of assessment method identified earlier and summarized in Table 1.
[b] Mean correlations are computed using Fisher's *r*-to-*z* transformation.
[c] The standard errors reported are for the standardized scores and are computed by S.E.$_{z_r} = 1/\sqrt{n-3}$.
[d] Gosling et al. (2003a, 2003b) used scales, not individual items, as indicators.
[e] The weighted mean of the minimum correlation for inter-observer agreement are calculated using only one of the two scores from Gosling et al. (2003a,b).
[f] Stephen and Ledger (2003) used Spearman's rank test, and thus reported correlations as rho.
[g] This study included a 75-item questionnaire, which was analyzed to reveal five temperament dimensions.
[h] We have not calculated the unweighted and sample-weighted mean minimum test–retest reliability because the mean would be based on only one correlation.

Table 4
How reliable are personality measures of dogs?—internal consistency

Study	Internal consistency of factors					Sample size	Total number of items in study[a]
	Mean α	Maximum		Minimum			
		α	Item label	α	Item label		
Gosling et al. (2003a) owner judgments of own dog	.83	.89	Neuroticism	.77	Extraversion	78	43
Gosling et al. (2003a) peer judgments of dog	.82	.86	Neuroticism	.75	Openness	78	43
Hsu and Serpell (2003)	.81	.93	Stranger-directed, Aggression	.67	Pain sensitivity	2054	132
Serpell and Hsu (2001)	.65	.84	Stranger-directed, fear/aggression	.53	Attachment	1067	38
Seksel et al. (1999)	.56	.73	Novel stimuli	.42	Handling	60	21
Weighted mean	.76	.90		.62			

Note: All Cronbach's alphas are as reported by the authors, not standardized.
[a] All of the studies except Gosling et al. (2003a) reported dropping items for various reasons.

these studies, by Goddard and Beilharz (1986), reveals Activity level is reliable from test to test, but that this reliability decreases as puppies age. The other study, by Netto and Planta (1997), shows a strong mean correlation, but also included many insignificant correlations. Closer examination reveals that many of the Kappa coefficients reported are zero, indicating no reliability. However, this is partially an artifact of the testing situation because the subtests were not intended to elicit Aggression, so it makes little sense to assess the reliability with which they elicited aggression. Of the subtests in this study which were intended to elicit aggression, the lowest Kappa coefficient is −.03 for reaction to an artificial hand taking away food, and reaction to a stranger being mildly threatening when meeting the dog's handler. However, Netto and Planta's study should be commended for fully reporting their reliability data; when interpreted against an understanding of the testing situations, these are data are very valuable.

Table 4 summarizes all the internal consistency estimates reported in the studies reviewed. Internal consistency measures estimate the degree to which items on a scale assess the same construct. In human personality research, they are often used following factor analyses to determine the internal coherence of the derived factors. Of the 16 studies in our review to focus on factor analysis, only three reported internal consistency. Two of these studies (Hsu and Serpell, 2003; Serpell and Hsu, 2001) gathered data using questionnaires with 5-point frequency (Likert) scales; the third (Seksel et al., 1999) used a 100-point scale. One additional study that did not focus on factor analysis also reported internal consistency (Gosling et al., 2003a) and is included in Table 4.

Internal consistency varied greatly across studies and factors, ranging from .42 for "Handling," to .93 for "Stranger-directed Aggression." Although high consistency is possible, it is by no means guaranteed. Nonetheless, the internal consistency measures had a weighted mean of .76, well within the limits acceptable in most human personality research (John and Benet-Martinez, 2000).

6.1. Summary

As a whole, the review of reported reliability coefficients is both encouraging and disappointing. The findings are encouraging because they show it is possible to measure dog temperament reliably using a variety of assessment methods. The findings are disappointing because they show that very few articles report reliability scores. Fortunately, there is an easy remedy—future researcher should compute and report the reliability of their measures.

7. Are assessments of dog temperament valid?

Once the reliability of a test has been established, the next step is to evaluate its construct validity. Validity is an index of how well an instrument is measuring what it is designed to measure. The construct validation process involves determining how well a measure assesses a construct (e.g., Fearfulness) as that construct has been conceptualized. A full conceptualization of a construct involves specifying the things to which the construct should be related and also the things to which the construct should be unrelated (Cronbach

and Meehl, 1955). These two components are known as convergent and discriminant validity. Convergent validity is supported when a measure correlates with other measures to which it should be related. Discriminant validity is supported when a measure is empirically unrelated to other measures that are theoretically unrelated (Campbell and Fiske, 1959). Thus, for example, the construct validity of a measure of Fearfulness would be supported by strong correlations with other measures of Fearfulness (i.e., convergent validity) and weak correlations with measures of theoretically unrelated traits, such as Sociability (i.e., discriminant validity; Devellis, 1991). To evaluate the validity of the tests in our review, we culled all potentially relevant validity data from the articles.

7.1. Obtaining and categorizing the validity coefficients

Our goal was to summarize the validity findings for each the seven broad temperament dimensions identified above (Reactivity, Fearfulness, Activity, Sociability, Responsiveness to Training, Submissiveness, and Aggression). Given these meta-analytic goals, we had to exclude from our analyses validity indices that were unique or could not be compared with other indices. For example, although Serpell and Hsu (2001); Hsu and Serpell, 2003) addressed validity directly they report only the significance levels of the Mann–Whitney U-statistics, not effect sizes, so their results could not be combined with the 11 other studies reporting validity, all of which report effect sizes.

Most studies did not explicitly conceptualize their findings in terms of convergent and discriminant validity and even those that did assess convergent validity or discriminant validity typically did not employ these terms. Therefore, after identifying all the potentially relevant validity coefficients, we devised a procedure for dividing them into three categories: convergent correlations, discriminant correlations, and indeterminate correlations. In studies where clear predictions were made, we could easily classify the correlations. Specifically, where a trait was predicted to correlate with a behavior, the resulting correlation was considered as evidence for convergent validity, and where a trait was predicted not to correlate with a behavior, the resulting correlation was considered evidence for discriminant validity.

However, when studies examined relationships between assessment scores and future behavior or future assessments but did not make predictions about these relationships, we needed a systematic procedure for assigning the correlations to the convergent, discriminant, or indeterminate categories. Thus, for each of these studies, we extracted descriptions of (1) the predictor variables (the trait or behavior assessed and how it was assessed), and (2) the validity criteria (the future behavior or later assessment results). Two expert judges who were blind to the actual findings of these studies made judgments about whether the predictor–criterion pairs should theoretically be related or unrelated. The two judges first made their judgments independently, then compared their judgments and discussed points of disagreement until consensus was reached. Those correlations associated with predictor–criterion matches were assigned to the convergent validity category and those correlations associated with predictor–criterion mismatches were assigned to the discriminant validity category. For example, the predictor–criterion pair in which adult dogs' wariness was a predictor of later problem behavior (Hennessy et al., 2001) was assigned to the convergent validity category, and the predictor–criterion pair in which the number of

objects a puppy investigated in a given period of time was a predictor of the adult dog's ability to cooperate (Wilsson and Sundgren, 1998) was assigned to the discriminant validity category. Of course, it should be noted that despite our best efforts to be comprehensive and systematic, the validity coefficients we report are inevitably influenced by our procedures for selecting coefficients and our findings should be evaluated in this light.

7.2. Convergent validity

Table 5 summarizes the available evidence for convergent validity. The convergent validity coefficients are organized in terms of the seven temperament dimensions plus two additional broader dimensions (problem behavior and broader evaluations of temperament), which are listed in the first column of the table. The second column lists the relevant citation. The third column lists the trait evaluated, as it was labeled by the original authors. The fourth column lists the criterion against which the trait was evaluated. The fifth column briefly summarizes the procedures by which the criterion data were obtained. The sixth column provides the validity coefficient as Pearson correlations or Spearman's rho. The seventh and eighth columns list the lower and upper bounds of the 95% confidence intervals for the unweighted mean estimates at the broad dimension level. The final column lists the sample size on which the validity coefficient was based.

Thus, the table shows, for example, that Ledger and Baxter (1996) examined the validity of Excitability ratings of 40 dogs made by their owners after adoption. The criterion by which the Excitability ratings were evaluated was behavior in response to an unfamiliar tester entering the dog's kennel. The sixth column shows that the owners' Excitability ratings correlated .64 with the dogs' behaviors when a stranger entered the kennel.

The summary statistics presented in Table 5 include both unweighted and a sample-weighted means. Both estimates are included because the sample sizes varied substantially across studies. For example, studies reporting convergent validity data on Reactivity had sample sizes ranging from 9 (Weiss and Greenberg, 1997) to 630 (Wilsson and Sundgren, 1998). The study of nine dogs reported a Reactivity convergent validity estimate of .36, whereas the study of 630 dogs reported Reactivity convergent validity estimates of .01 and .05. The mean validity coefficient for Reactivity is .35 if averaged across all studies but .09 if weighted by the number of dogs in each study. Both estimates are potentially interesting, with the first estimate giving equal weight to each study and the second estimate giving equal weight to each individual dog tested.

Overall, the evidence for convergent validity is reasonably promising, with the various estimates averaging about .40 across the nine dimensions examined here. However, the findings do show some variability across the dimensions. The dimensions with the fewest studies will tend to provide the least stable estimates so it is not surprising that highest and lowest validity estimates are associated with the dimensions with very studies. In particular, the strongest convergent validity coefficients (unweighted mean = .88, sample-weighted mean = .88) are associated with the Submissiveness dimension. However, with rather divergent evidence from only two studies, the confidence intervals around this mean are enormous, ranging from 0 to 1. Therefore, we do not feel confident providing a validity estimate for this dimension. Clearly, more research is needed before estimates can be made about the validity of Submissiveness assessments.

Table 5
Convergent validity: how well do dog temperament tests predict future behavior or scores on other assessments?

Dimension study	Trait	Criterion measure		Validity coefficient	95% Confidence interval		# of subjects
		Criterion behavior	Basis for scoring		Lower	Upper	
Reactivity							
Ledger and Baxter (1996) [a]	Excitability (rated by owners after adoption)	Un-named	Unspecified behavioral response to an unfamiliar tester in kennel	.64			40
Stephen and Ledger (2003) [a]	Behavior problems towards strangers (rated by owners after adoption)	Excitability towards tester	Behavior when tester greets/meets the dog	.32			40
			Unspecified behavioral response to grooming	.66			40
Goddard and Beilharz (1986) [b]	Excitability score (rated by trainers)	Composite of scores on sit, activity tests	Repetitions dog needs to "sit" on command; number of movements	.22			102[c]
Weiss and Greenberg (1997) [a]	Excitement (rated by three observers)	Excitement-related behaviors	Scoring method not specified, but behaviors included steady high level of jumping, pawing, barking, etc.	.36			9
Wilsson and Sundgren (1998) [d]	Prey drive[e] (rated by trainers)	Fetching	Time until puppy picks up tossed ball	.01[f]			630
		Retrieving	Willingness scored with set criteria	.05			630
Unweighted mean				.35	.07	.57	
Sample-weighted mean				.09			
Fearfulness							
Hennessy et al. (2001; puppies)[g,h]	Part of overall problem index (rated by owners after adoption)	Flight	Number of movements to escape; time in door well, jumps	.02/.34[j,k]			23/18[l]

Table 5 (*Continued*)

Dimension study	Trait	Criterion measure		Validity coefficient	95% Confidence interval		# of subjects
		Criterion behavior	Basis for scoring		Lower	Upper	
Hennessy et al. (2001; adults)[g,i]	Part of overall problem index (rated by owners after adoption)	Flight	Number of attempts to escape, time spent in door well, jumps	.47/.74			10/7[l]
Hennessy et al. (2001; puppies)[g,h]	Part of overall problem index (rated by owners after adoption)	*Timidity*	Time spent in door well	.39[j,k]/.11			23/18[l]
Hennessy et al. (2001; adults)[g,i]	Part of overall problem index (rated by owners after adoption)	*Timidity*	Time spent in door well	.03/.37			10/7[l]
Hennessy et al. (2001; puppies)[g,h]	Part of overall problem index (rated by owners after adoption)	Wariness	Latency to contact toy car, horn	.43[j]/.31[k]			23/18[l]
Ledger and Baxter (1996) [a]	Timidity	Un-named (rated by owners after adoption)	Unspecified behavioral response to being walked on-leash;	.68			40
			Unspecified behavioral response to being approached by a person with a ''titbit''	.79			40
Stephen and Ledger (2003) [a]	Fearfulness (rated by owners after adoption)	Tester observations through	out test reported as ''not correlated''				
Gosling et al. (2003a)	Extraversion	Extraversion-related behavior (rated by owner)	Observer rating based on a variety of field-test behaviors, during greetings, etc.	.32			78
Goddard and Beilharz (1984a)	General Nervousness (rated by trainers)	Fear on walk (3 months)	Ratings by trainers based on a combination of reactions to various stimuli, including clap noise, toy horse, gun shot, party whistle, rapid head movement, ear position,stranger entering house	.24			102[c]
	General Nervousness (rated by trainers)	Fear on walk (4 months)		.35			102[c]

	General Nervousness (rated by trainers)	Fear on walk (6 months)		.42			102[c]
	General Nervousness (rated by trainers)	Fear on walk (12 months)		.58			102[c]
	General Nervousness (rated by trainers)	Fear on walk (day 3 of final evaluation)		.59			102[c]
	General Nervousness (rated by trainers)	Fear on walk (day 4 of final evaluation)		.64			102[c]
Goddard and Beilharz (1986) [b]	General Fearfulness (rated by trainers)	Composite of fear on walk, Reactions to specific stimuli at different ages	Fear on walk—see Goddard and Beilharz (1984a); other tests include reaction to surfboard at 10 weeks, activity level during Handling at 9 weeks, etc.	.57			102[c]
Unweighted mean				.45	.33	.55	
Sample-weighted mean				.48			
Activity							
Hennessy et al. (2001; adults) [g,i]	Part of overall problem index (rated by owners after adoption)	Locomotor activity	Number of movements to withdraw, approach, line crossing	.05[k]/.16[j]			10/7[l]
Beaudet et al. (1994)	Activity level (retested using same assessment at 1.61 and 3.68 months)	Locomotor activity	Number of line crossings in test chamber	.04[m]			39
Unweighted mean				.08	0[n]	.25	
Sample-weighted mean				.06			
Sociability							
Hennessy et al. (2001; adults) [g,i]	Part of overall problem index (rated by owners after adoption)	Sociability	Time spent in proximity to a person, stationary, latency to contact person, time in door well, person contact, etc.	.21/.63			10/7[l]
Hennessy et al. (2001; puppies) [g,h]	Part of overall problem index (rated by owners after adoption)	*Timidity*	Time spent in door well (includes avoiding people)	.39[i,k]/.11			23/18[l]

Table 5 (*Continued*)

Dimension study	Trait	Criterion measure		Validity coefficient	95% Confidence interval		# of subjects
		Criterion behavior	Basis for scoring		Lower	Upper	
Hennessy et al. (2001; adults) [g,i]	Part of overall problem index (rated by owners after adoption)	*Timidity*	Time spent in door well (includes avoiding people)	.03/.37			10/7[l]
Stephen and Ledger (2003) [a]	Playfulness (rated by owners after adoption)	Play behaviors with tester	Play displays during tug-o-war with tester grooming by tester having lead put on by tester being walked on lead by tester	.53			40
				.44			40
				.33			40
				.49			40
Gosling et al. (2003a)	Neuroticism (rated by owners)	Neuroticism-related behavior	Observer rating based on a variety of field-test behaviors, during greetings, etc.	.21			78
Wilsson and Sundgren (1998) [d]	Affability (rated by trainers)	Yelping	Time until puppy (alone) whines/yelps	.00[l,k,o]			630
Unweighted mean				.34	.19	.45	
Sample-weighted mean				.12			
Responsiveness to Training							
Ledger and Baxter (1996) [a]	Obedience (rated by owners after adoption)	Un-named	Unspecified behavioral response to showing dog its leash, saying ''walkies''	.72			40
Stephen and Ledger (2003) [a]	Obedience (rated by owners after adoption)		Tester observations through-out test	Reported as ''not correlated''			
van der Borg et al. (1991)	Disobedience (rated by owners after adoption)	Disobedience	Questionnaire to care-taker	.41[p]			81
			Testers' evaluation of disobedience	.27[p]			81

van der Borg et al. (1991)	Pulling on leash (rated by owners after adoption)	Pulling on leash	Questionnaire to care-taker	.51[p]			81
			Testers' evaluation of pulling	.16[p]			81
Gosling et al. (2003a)	Openness	Openness-related behavior (rated by owner)	Observer rating based on a variety of field-test behaviors, during greetings, etc.	.23			78
Weiss and Greenberg (1997) [a]	Attention/distraction (rated by three observers)	Attention/distraction-related behaviors	Scoring method not specified, but behavior described: ''dog's attention should be on the handler''	.00			9
Wilsson and Sundgren (1998) [d]	Ability to Cooperate (rated by trainers)	Contact	Reaction to, attempt to contact person	.17[f,k]			630
Unweighted mean				.33	.10	.52	
Sample-weighted mean				.25	.25		
Submissiveness							
Weiss and Greenberg (1997) [a]	Dominance (rated by three observers)	Dominance-related behaviors	Scoring method not specified, but behaviors included front paws on handler, mounting, placing body above handler, growling while making eye contact	.13			9
Weiss and Greenberg (1997) [a]	Fear/Submission (rated by three observers)	Fear/Submission-related	Scoring method not specified, but behaviors included crouching, submissive urination, shoulder roll, prolonged startle/fear to noise, etc.	1.00[q]			9
Unweighted mean				.88	0[n]	1	
Sample-weighted mean				.88			
Aggression							
Ledger and Baxter (1996) [a]	Aggression (rated by owners after adoption)	Un-named	Unspecified behavioral response to showing dog its leash, saying ''walkies'';	.82			40
			Unspecified behavioral response to playing tug-o-war	.82			40

Table 5 (*Continued*)

Dimension study	Trait	Criterion measure		Validity coefficient	95% Confidence interval		# of subjects
		Criterion behavior	Basis for scoring		Lower	Upper	
van der Borg et al. (1991)	Aggression towards (rated by owners after adoption)	Aggression towards adults	Questionnaire to care-taker	.45[p]			81
			Testers' evaluation of aggression	.26[p]			81
van der Borg et al. (1991)	Aggression towards dogs (rated by owners after adoption)	Aggression towards dogs	Questionnaire to care-taker	.55[p]			81
			Testers' evaluation of dog-related aggression	.23[p]			81
Gosling et al. (2003a)	Agreeableness (rated by owner)	Agreeableness-related behavior	Observer rating based on a variety of field-test behaviors, during greetings, etc.	.33			78
Netto and Planta (1997)	Bite history	Aggression, tendency to bite (reported by owner)	Observed biting attempts and snapping during 43 subtests of Test Battery	.25[j]			112
Netto and Planta (1997)	Bite history	Aggression, tendency to bite (reported by owner)	Observed biting attempts (without snapping) during 43 subtests of Test Battery	.31[r]			112
Wilsson and Sundgren (1998) [d]	Prey drive[e] (rated by trainers)	Fetching	Time until puppy picks up tossed ball	.01[f,k]			630
		Retrieving	Willingness scored by set criteria	.05			630
Unweighted mean				.42	.18	.61	
Sample-weighted mean				.18			
Unweighted mean across all seven temperament dimensions				.41			
Sample-weighted mean across all seven temperament dimensions				.22			
Problem behaviors							

Hennessy et al. (2001; puppies)[g,h]	Part of overall problem index (rated by owners after adoption)	Solicitation	Number of jumps against observation platform	.09/.18			23/18[l]
Hennessy et al. (2001; adults)[g,i]	Part of overall problem index (rated by owners after adoption)	Solicitation	Number of jumps against observation platform	.54/.72			10/7[l]
Ledger and Baxter (1996) [a]	Separation-related problems (rated by owners after adoption)	Un-named	Unspecified behavioral response to being approached by a person with a ''titbit''	.82			40
van der Borg et al. (1991)	Car-related problems (rated by owners after adoption)	Car-related problems	Questionnaire to care-taker	.20[p]			81
			Testers' evaluation of car-related problems	.23[p]			81
van der Borg et al. (1991)	Separation anxiety (rated by owners after adoption)	Separation anxiety	Questionnaire to care-taker	.66[p]			81
			Testers' evaluation of separation problems	.22[p]			81
Unweighted mean				.45	.19	.66	
Sample-weighted mean				.41			
Broad evaluation of temperament							
Weiss and Greenberg (1997) [a]	Completion of a set of tasks in final test	General selection test	Scored by tester on various tasks and subjective ''feeling''	.18			9
Weiss and Greenberg (1997) [a]	Number of corrections needed to complete tasks in final test	General selection test	Scored by tester on various tasks and subjective ''feeling''	.21			9
Beaudet et al. (1994)	Cumulative Social Tendency Score (Submissiveness/Dominance) (retested using same assessment at 1.61 and 3.68 months)	Social attraction	Puppy's reaction during 30 s of tester crouching, coaxing puppy to the tester	.29			39
		Following	Puppy's reaction when tester tries to coax puppy to walk by the tester				

Table 5 (*Continued*)

Dimension study	Trait	Criterion measure		Validity coefficient	95% Confidence interval		# of subjects
		Criterion behavior	Basis for scoring		Lower	Upper	
		Restraint dominance	Puppy's reaction when tester holds puppy n its back for 30 s				
		Elevation dominance	Puppy's reaction when tester holds puppy 15 cm off the ground for 30 s				
		Social dominance	Puppy's reaction to being stroked from head to tail for 30 s				
Beaudet et al. (1994)	Cumulative Social Tendency Score	Locomotor activity at 1.61 months(Submissiveness/ Dominance) (tested at 1.61 months)	Number of line crossings in test chamber	.45			39
Beaudet et al. (1994)	Cumulative Social Tendency Score	Locomotor activity at 3.68 months (Submissiveness/ Dominance) (tested at 3.68 months)	Number of line crossings in test chamber	.70			39
Unweighted mean				.39	.06	.64	
Sample-weighted mean				.46			
Unweighted mean across all dimensions, including problem behaviors and broad evaluations of temperament				.41			
Sample-weighted mean across all dimensions, including problem behaviors and broad evaluations of temperament				.24			

Notes: Mean correlations are computed using Fisher's *r*-to-*z* transformation.

[a] These correlations are rho values from Spearman's rank analysis.

[b] Goddard and Beilharz (1986) report extensively on the correlations between components to these overall scores and the trait they were used to predict. We have not reported all of these coefficients individually because they are components to the overall scores and to do so would skew our overall correlations. Please see the original source for more details of these component correlation coefficients.

[c] Goddard and Beilharz (1984a, 1986) reported an original *N* of 102 before an unspecified number of subjects that dropped out.

[d] Wilsson and Sundgren (1998) examined all possible correlations but reported effect sizes for significant correlations only.

[e] In our sorting procedure, the behavior of Prey drive was categorized into both Aggression and Reactivity and is thus listed twice here.

[f] The correlations between Yelping and Affability, Contact and Ability to Cooperate, and Fetching and Prey drive were all reported as negative such that a shorter latency (less time) to Yelping correlates with higher Affability, a shorter latency to make contact correlates with a higher adult score on Ability to Cooperate, and a shorter latency to pick up a thrown ball correlates with a higher adult Prey drive. These correlations have been rekeyed so that a higher correlation reflects greater validity.

[g] Hennessy et al. (2001) received so few reports of problem behaviors that it was deemed necessary to create a combined ''behavior problems'' score instead of attempting to examine prediction of individual types of behavior problems.

[h] These assessments were performed with puppies or juvenile dogs who still have their milk teeth.

[i] These assessments were performed with juvenile or adult dogs, or dogs who have their adult teeth.

[j] These correlations were all reported as negative such that, for example, a higher level of Locomotor Activity as a puppy correlated with fewer behavior problems as an adult. These correlations have been rekeyed so that a higher correlation reflects greater validity.

[k] These correlations are opposite what was predicted (e.g., a positive correlation was expected, but a negative was found).

[l] Owners were asked to rate their new pets 2 weeks after adoption, and then at 6 months after adoption. The *N*'s 2 weeks after adoption are larger than 6 months later for both puppies and juvenile/adult dogs.

[m] The correlation between number of movements at 1.61 and 3.68 months was reported as negative (but not significant).

[n] We have truncated these confidence intervals to reflect the range of possible convergent-validity coefficients. Calculation of the intervals from the correlations provided yields confidence intervals ranging from less than zero, which is clearly impossible when addressing validity.

[o] Due to rounding, this correlation is reported as 0, but it is actually .001 and significant.

[p] We have calculated the validity coefficients for van der Borg et al. (1991) from the data the authors provided.

[q] Calculations involving the reported $r = 1$ are calculated using $r = .99$; when $r = 1$, Fisher's r-to-z yields a z of infinity, because a true correlation of $r = 1$ occurs with the probability of 0.

[r] We have calculated the validity coefficients for Netto and Planta (1997) from the data the authors provided. Netto and Planta (1997) also report 15.4% false positives, or that 15.4% of the dogs they predicted from their test to have a bite history do not/have never bitten before.

The lowest validity coefficients (unweighted mean = .08 with a 0–.25 confidence interval; sample-weighted mean = .06) are associated with the Activity dimension. However, only two studies report convergent validity coefficients associated with this dimension, for a total of three coefficients, again suggesting the need for further research.

The strongest interpretable validity coefficients (unweighted mean = .48, sample-weighted mean = .45) are associated with the Fearfulness dimension. Fearfulness was examined in many studies and with many different predictors. It may not be surprising that Fearfulness is associated with strong validity coefficients and a relatively narrow confidence interval (.33–.55), because this dimension has been shown to be relatively highly predictable, even from early puppyhood to later adulthood (e.g., Goddard and Beilharz, 1984b).

What criteria should be used to evaluate these validity coefficients? One potential benchmark is provided by equivalent research in the human literature. In human studies, trait-behavior correlations are typically in the order of .20–.30. For example, in one human study, correlations between self-reported personality and ratings made by observers after a 20-min discussion task averaged .24 across the Big Five human personality dimensions (Paulhus and Bruce, 1992). Measured against this human standard, the dog validity coefficients seem satisfactory at the very least.

As shown in Table 5, the convergent validity coefficients varied substantially across the studies, with some studies obtaining much stronger validity estimates than others. What factors could be driving the cross-study differences in validity? One possibility is the age of the dogs. Indeed, indirect support for the idea that puppies are harder to test than older dogs is provided by the fact that the study with the lowest average validity coefficient (less than .05) involved puppies (Wilsson and Sundgren, 1998). More generally, there is a marked difference between the validity coefficients for tests of puppies (sample-weighted mean $r = .14$; Beaudet et al., 1994; Hennessy et al., 2001; Goddard and Beilharz, 1984a, 1986; Wilsson and Sundgren, 1998) versus adult dogs (sample-weighted mean $r = .43$).

Two of the studies provide more direct support for this idea. Hennessy et al. (2001) evaluated the validity of assessments administered in the same way both to puppies and to older dogs; the mean validity coefficients for the puppies (unweighted mean = .25; sample-weighted mean = .25) was much lower than that for the older dogs (unweighted mean = .41; sample-weighted mean = .37). This is consistent with another study, which identified a nearly linear relationship between age and test validity (Goddard and Beilharz, 1984a). Together these studies strongly suggest that tests of young puppies are relatively poor predictors of their future behavior compared to tests of older dogs. These tests suggest that the inclusion of puppy studies in our meta-analysis biases the estimate of validity in older dogs. Indeed, if we remove the results of the large study of 630 puppies (Wilsson and Sundgren, 1998) from the meta-analysis, the overall sample-weighted validity estimate assessed across all seven temperament dimensions jumps from .23 to .42.

7.3. Discriminant validity

Although discriminant validity has largely been neglected, two articles (Hsu and Serpell, 2003; Serpell and Hsu, 2001) did examine and report this facet of construct

validity. Overall, Hsu and Serpell found good evidence for the discriminant validity of the measures, although there were a few exceptions (e.g., an unpredicted association between attachment and stranger fear/stranger aggression). However, even these exceptions are useful because they can serve as a launching point for future studies that investigate these unexpected links.

In addition to the Serpell and Hsu studies, there were some other studies that mentioned discriminant validity but did not report the relevant correlations (Goddard and Beilharz, 1986; van der Borg et al., 1991) and there were some studies that reported the relevant correlations although they did not describe them in terms of discriminant validity (Hennessy et al., 2001; Wilsson and Sundgren, 1998). These latter studies were identified by the procedures described in Section 7.1 above.

In Hennessy et al. (2001) study of temperament in shelter animals, our validity-categorization procedures identified six potential discriminant correlations. For example, this study reported correlations between puppies' locomotor activity and the conceptually unrelated incidence of problem behavior measured two weeks ($r = -.25$) and 6 months ($r = -.30$) after adoption. The absolute values of the discriminant correlations averaged .37 across the six estimates. Although none of these values were significant, these values were no lower than the convergent correlations from the same study (which averaged .36). This pattern of findings did not match the pattern of findings required to support discriminant validity, in which the convergent correlations should substantially exceed the discriminant correlations. Thus, there was no support for the discriminant validity of these assessments in this study.

Wilsson and Sundgren (1998) computed a very large number of validity correlations but reported the effect sizes only for those correlations that were statistically significant. However, because their sample size was very large, even very small coefficients reached significance. Indeed, the one statistically significant discriminant validity estimate (between the number of objects puppies visited when placed in a room containing novel objects and adult defense drive) had a very small effect size (.024). Of course, the numerous discriminant validity correlations that did not reach statistical significance can also reasonably be taken as evidence for the discriminant validity of the corresponding measures (because these measures also exhibited convergent validity). Unfortunately, however, as in the Hennessy et al. (2001) study, the convergent correlations in the Wilsson and Sundgren study did not substantially exceed the discriminant correlations.

7.4. Summary and discussion of validity findings

Taken as a whole, the evidence broadly supports the convergent validity of temperament assessments in dogs, especially adult dogs, but there was only mixed evidence for discriminant validity. However, these conclusions are based on a rather small proportion of the literature as most studies did not address validity issues. Given the centrality of validity in any assessment context, further examination of validity should remain a top priority for dog temperament researchers. In particular, research is needed to establish the parameters (e.g., dog age, testing context) that could affect validity; such findings will be essential for future work in both research and applied contexts.

In addition to furnishing numerical estimates of validity, our review of the validity literature revealed a couple of notable trends. First, an unusually large proportion (90%) of the validity studies were based on Ratings of Individual Dogs; this should be contrasted with the fact that Ratings of Individual Dogs are relatively rare (17%) compared with the other methods of assessment. Second, although studies of shelter dogs constitute a small proportion of the studies in our review (14%), they were assessed in half the studies of validity. It would seem that researchers working in shelter contexts are particularly concerned with measurement issues; indeed, five of the seven shelter-dog studies reported the validity of their temperament tests, and, of the two that did not, one focused on the reliability of temperament testing.

We conclude by noting a trend that pervades temperament and personality research on other species (Gosling et al., 2003b). We highlight it here because although it is typically missed or ignored, it has substantial implications for validity. Research on the reliability of the measures of the criterion behaviors (against which the ratings are tested) is almost non-existent; the reliability of behavioral codings such as the number of movements to escape (see Table 5) is often assumed but is rarely tested. We suspect that researchers assume that the reliability of behavioral codings will be high because such codings seem objective. That is, behavioral codings like the number of movements a puppy makes in a given time period (Hennessy et al., 2001) appear more objective than do ratings of temperament, but research on humans has shown substantial variability in the reliability of such behavioral codings (Gosling et al., 1998). Therefore, it is essential that future validity research should assess and report the reliability of the criterion measures against which the validity of other means of assessment are to be estimated. Without this information, it is impossible to know whether low validity correlations are low due to genuinely low validity or due to the low reliability of the criterion measures.

8. Summary and conclusions

By bringing together the disparate research on temperament in dogs, our review allowed us to summarize what is known about canine temperament and to identify some trends and gaps in the field. Below we summarize our conclusions and, where appropriate, we highlight some directions for future research.

(1) An extensive literature search identified 51 empirical publications on dog personality or temperament. The articles, published between 1934 and 2004, varied greatly in their assessment methods, research goals, and the attributes of their subjects (in terms of breed, age, breeding and rearing environment, and sexual status). In addition, the studies also varied in their methodological rigor, with some studies being little more than a few informal observations of a handful of dogs and others being large-scale systematic multi-phase assessments.
(2) We found that dog temperament assessment methods can be usefully grouped into four categories, which we have called Test Batteries, Ratings of Individual Dogs, Expert Ratings of Breed Prototypes, and Observational Tests. A fifth category represents studies that combined more than one assessment method. The most

common assessment method was the Test Battery, which was, in theory, the closest of the four methods to achieving objectivity. In practice, however, the levels of objectivity attained differed considerably. Future research should focus on direct comparisons of the methods in terms of reliability, validity, and efficiency in different research contexts.

(3) Our review showed that dog temperament studies varied in their research goals (e.g., examining behavioral tendencies specific to breeds, Hart and Miller, 1985; Mahut, 1958; Svartberg and Forkman, 2002; predicting adult police dog performance from puppy behavior, Slabbert and Odendaal, 1999). The vast majority of dogs tested were in working contexts (e.g., as guide or police dogs), with a relatively small number of pet or shelter dogs being studied. Given the high demand for temperament testing in shelters and to assess whether dogs are fit to be adopted, greater research attention should be directed towards pet and shelter dogs. And until studies have been done to establish the generalizability of findings from working dogs to pet dogs, generalizations from one population to another should be made with caution.

(4) In the studies in our review reporting breed, at least 85% of the dogs were purebred. The Labrador Retriever and the GSD were the most frequently represented breeds, combining to compose 32% of the subjects. The GSD was by far the most frequently tested breed, composing 26% of the dogs tested (9205 dogs). A small minority of dogs were the planned offspring of two purebreds of different breeds, and there were also very few dogs of unintentional or unknown breed mixtures. Although this makes sense insofar as the Labrador and the GSD are two of the most frequently registered breeds in the AKC, little work has been done to examine the generalizability of these findings to different breeds. One of the few studies to compare temperament across breed examined large populations of both of these commonly assessed breeds, the GSD and the Labrador Retriever, and found substantial differences in temperament (Wilsson and Sundgren, 1997). Another study to examine differences among groups of breeds (e.g., Terriers, Scent hounds, Sheepdogs, etc.) again found significant differences among the breeds, indicating that some breed groups display unique patterns of temperament (Svartberg and Forkman, 2002). Unfortunately, a substantial number of studies failed to report breed information. By neglecting to examine breed as a potentially important influence on temperament, the value of such studies is diminished. Future research should concern itself with gaining a fuller representation of dog breeds and with providing breed information, further elucidating breed- and breed group-specific temperament patterns.

(5) We also found that some method-breed combinations are more common than others. About one third of the dogs in Test Battery studies are GSDs being tested for their potential as police and working dogs. Eighty percent of all dogs in studies using Observational Tests are Labrador Retrievers, tested for their potential as guide dogs. Future research should examine the effectiveness of these two test methods, particularly Observational Tests, for other breeds and other purposes, because their ability to generalizability beyond such specific contexts cannot be assumed.

(6) There is also an age-related bias in the studies. Most studies examine dogs who were young or still in puppyhood when tested, and only few studies looked at dogs over the age of four years. In addition, age effects were rarely examined in studies using

Ratings of Individual Dogs and Expert Ratings of Breeds. Consequently, we know little about how aging may shape temperament in dogs. Future research should focus on this question, and examine the developmental trajectory of temperament in dogs. In particular, future research should identify the point at which temperament stabilizes, such that adult traits can be predicted from puppy behavior.

(7) Eighteen of the studies in our review examined dogs bred for particular programs. Some of these studies used scores on temperament tests as the basis for selective breeding. After several generations, such selective-breeding programs may shape responses to temperament tests. Indeed, in one study, selective breeding lead to an increase in puppy test scores over successive generations, but the rates at which adult dogs became successful guide dogs did not match this increase (Scott and Bielfelt, 1976).

(8) Although most pet and shelter dogs are spayed or neutered, the vast majority of dogs assessed were intact. The rare studies that did examine the effects of castration indicated that intact male dogs were the most likely to show aggressive behavior, and intact female dogs were the least likely (Podberscek and Serpell, 1996; Roll and Unshelm, 1997). However, Podberscek and Serpell's study also revealed that neutering a dog in reaction to his aggressive behavior does not reduce future aggression. Obviously, given that aggressive behavior is a concern in many programs and to many private dog owners, additional systematic research is needed in this area.

(9) A systematic multi-step procedure for summarizing the traits that have been examined in previous canine research identified seven broad temperament dimensions: Reactivity, Fearfulness, Activity, Sociability, Responsiveness to Training, Submissiveness, and Aggression. Our sorting procedures revealed very little standardization in the terms used to describe dog temperament. Different studies often used the same terms to refer to different behaviors and different terms were often used to refer to very similar behaviors. There is clearly a need to develop a common language with which to describe temperament traits in dogs (Goodloe and Borchelt, 1998). We propose that the seven categories derived from our review of the literature represents a sensible starting point for developing such a standard lexicon of canine-temperament descriptors.

(10) The most frequently examined temperament dimension was Fearfulness, with traits related to this dimension appearing in 43 studies. Traits in the Fearfulness dimension were frequently also categorized in the Reactivity dimension, suggesting some conceptual and empirical overlap between these two dimensions. Further research on the traits of Reactivity and Fearfulness in dogs is needed to determine whether the two can be usefully distinguished or are better considered as two facets of an even broader superordinate category.

(11) Sociability was also studied frequently, in 31 studies. The traits categorized under this dimensions were sometimes also categorized under the Responsiveness to Training dimension. We suggest this overlap may be driven by the fact that an interest in people is central to both Sociability and interest in training. Future research should examine the extent to which Sociability determines Responsiveness to Training, and how best to isolate Responsiveness to Training as a separate dimension.

(12) Numerous studies included traits related to Activity. Our review showed that level of Activity changes dramatically with age. However, there was also some evidence that

Activity can moderate the expression of other traits. Future research should directly examine this important possibility.

(13) The studies that reported reliability were encouraging, showing that it is possible to assess dog temperament reliably. However, these findings must be tempered by the fact that these conclusions are based on a lamentably small number of studies. We were shocked to discover that very few studies even report the reliability of the measures they used. Clearly, given the importance of reliability in all assessment contexts, future research should examine and report reliability.

(14) Taken as a whole, the evidence broadly supports the convergent validity of temperament assessments in dogs. However, this conclusion is based on a rather small proportion of the literature as most studies do not address validity issues. Given the centrality of validity in any assessment context, further examination of validity should remain a top priority for dog temperament researchers. In particular, research is needed to establish the parameters that affect validity; such findings will be essential for future work in research and applied contexts.

(15) Although the overall convergent validity findings were generally encouraging, our review suggests that tests of young puppies are not valid predictors of their future behavior. Given that puppy tests are widely used but their validity is rarely examined, this finding has huge implications for work in applied and research contexts. Future research is urgently needed to examine this possibility directly.

(16) Our review showed that unusually large proportions of the validity studies were based on Ratings of Individual Dogs (90%) and used shelter dogs (50%). It seems that researchers working in shelter contexts are particularly concerned with measurement issues. However, such basic issues should be of concern to all dog temperament researchers.

(17) Although rating methods (e.g., of ''Fearfulness'') were well represented in the studies examining reliability and validity, studies examining the reliability and validity of behavioral codings (e.g., number of time the dog scratches) are almost non-existent. The reliability and validity of codings is often assumed but rarely tested. However, research on humans has shown substantial variability in the reliability and validity of such codings (Gosling et al., 1998). Therefore, future research should also assess and report the reliability and validity of behavioral codings. This is important in the context of validity because behavioral codings are often used as the criterion against which ratings are evaluated; but if the criterion behavioral codings are not measured reliably the ratings would appear to have low validity irrespective of their true validity.

(18) Past validity research has focused on convergent validity and generally neglected discriminant validity. Overall, the reported discriminant validity results were mixed. If the construct validity of dog temperament measures is to be established, it is important that future research examine both types of validity.

Over the past 70 years great strides have been made in our understanding of personality and temperament in dogs. This review, based on the published empirical research over this period, generally supports the viability of assessing canine temperament. In addition, the review provides a roadmap specifying the major empirical questions that need to be addressed in the next generation of studies.

Acknowledgements

Preparation of this article was supported by an NSF fellowship to Amanda C. Jones and a fellowship from the Center for Advanced Study in the Behavioral Sciences to Samuel D. Gosling. We are indebted to Mark Venghaus, Lisa Wolverton, Scott Sample, Janice Patton, Margaret Johnson, Barbara Smuts, Camille Ward, Erica Bauer, Kelly Hall, and Diane Mollaghan for their expert rating of the dog temperament traits and their support. We are thankful to Kelly Hall, and Kerri Sharp for their help with data collection, and to Matt Jones and Greg Hixon for their statistical expertise. We also thank Rebecca Ledger, Jacqueline Stephen, James A. Serpell, and Yuying Hsu for their cooperation in helping us to cull reliability and validity statistics from their research.

References

Barrick, M.R., Mount, M.K., 1991. The Big Five personality dimensions and job performance: a meta-analysis. Pers. Psychol. 44, 1–26.

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. Charles C. Thomas, Springfield, IL, 161 pp.

Bogg, T., Roberts, B.W., 2004. Conscientiousness and health behaviors: a meta-analysis. Psychol. Bull. 130, 887–919.

Bradshaw, J.W.S., Goodwin, D., 1998. Determination of behavioural traits of pure-bred dogs using factor analysis and cluster analysis: a comparison of studies in the USA and UK. Res. Vet. Sci. 66 (1), 73–76.

Buss, A.H., 1995. Personality: Temperament, Social Behavior, and the Self. Allyn and Bacon, Needham Hghts, MA, 408 pp.

Campbell, W.E., 1972. A behavior test for puppy selection. Mod. Vet. Pract. 12, 29–33.

Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol. Bull. 56, 81–105.

Cattell, R.B., Korth, B., 1973. The isolation of temperament dimensions in dogs. Behav. Biol. 9, 15–30.

Cattell, R.B., Bolz, C.R., Korth, B., 1973. Behavioral types in purebred dogs objectively determined by taxonome. Behav. Genet. 3 (3), 205–216.

Coren, S., 1995. The Intelligence of Dogs: A Guide to the Thoughts, Emotions, and inner Lives of our Canine Companions. Free Press, New York, NY, 288 pp.

Coren, S., 1998. Why we Love the Dogs we do: How to Find a Dog that Matches Your Personality. Free Press, New York, NY, 320 pp.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychol. Bull. 52, 281–302.

Devellis, R.F., 1991. Scale Development. SAGE Publications, Newbury Park, CA, 121 pp.

Draper, T.W., 1995. Canine analogs of human personality factors. J. Gen. Psychol. 122 (3), 241–252.

Dugatkin, L.A., 2004. Principles of Animal Behavior. Norton, New York, NY, 596 pp.

Goddard, M.E., Beilharz, R.G., 1982–1983. Genetics of traits which determine the suitability of dogs as guide-dogs for the blind. Appl. Anim. Ethol. 9, 299–315.

Goddard, M.E., Beilharz, R.G., 1984a. A factor analysis of fearfulness in potential guide dogs. Appl. Anim. Behav. Sci. 12, 253–265.

Goddard, M.E., Beilharz, R.G., 1984b. The relationship of fearfulness to, and the effects of sex, age and experience on exploration and activity in dogs. Appl. Anim. Behav. Sci. 12, 267–278.

Goddard, M.E., Beilharz, R.G., 1985. Individual variation in agonistic behaviour in dogs. Anim. Behav. 33, 1338–1342.

Goddard, M.E., Beilharz, R.G., 1986. Early prediction of adult behavior in potential guide dogs. Appl. Anim. Behav. Sci. 15, 247–260.

Goldsmith, H., Buss, A., Plomin, R., Rothbart, M., Thomas, A., Chess, S., Hinde, R., McCall, R., 1987. Roundtable: what is temperament? Four approaches. Child Dev. 58, 505–529.

Goodloe, L.P., Borchelt, P.L., 1998. Companion dog temperament traits. J. Appl. Anim. Welfare Sci. 1 (4), 303–338.

Gosling, S.D., 2001. From mice to men: what can we learn about personality from animal research? Psychol. Bull. 127, 45–86.

Gosling, S.D., Bonnenburg, A.V., 1998. An integrative approach to personality research in anthrozoology: ratings of six species of pets and their owners. Anthrozoös 11 (3), 148–156.

Gosling, S.D., John, O.P., 1999. Personality dimensions in non-human animals: a cross-species review. Curr. Dir. Psychol. Sci. 8, 69–75.

Gosling, S.D., Kwan, V.S.Y., John, O.P., 2003a. A dog's got personality: a cross-species comparative approach to evaluating personality judgments. J. Pers. Soc. Psychol. 85, 1161–1169.

Gosling, S.D., Lilienfeld, S.O., Marino, L., 2003b. Personality. In: Maestripieri, D. (Ed.), Primate Psychology: The Mind and Behavior of Human and Nonhuman Primates. Harvard University Press, Cambridge, MA, pp. 254–288.

Gosling, S.D., John, O.P., Craik, K.H., Robins, R.W., 1998. Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. J. Pers. Soc. Psychol. 74, 1337–1349.

Hart, B.L., 1995. Analyzing breed and gender differences in behaviour. In: Serpell, J. (Ed.), The Domestic Dog: Its Evolution, Behaviour and Interactions with People. Cambridge University Press, Cambridge, England, pp. 65–78.

Hart, B.L., Hart, L.A., 1985. Selecting pet dogs on the basis of cluster analysis of breed behavior profiles and gender. JAVMA 186 (11), 1181–1185.

Hart, B.L., Miller, M.F., 1985. Behavioral profiles of dog breeds. JAVMA 186 (11), 1175–1180.

Hart, B.L., Murray, S.R.J., Hahs, M., Cruz, B., Miller, M.F., 1983. Breed-specific behavioral profiles of dogs: model for a quantitative analysis. In: Katcher, A.H., Beck, A.M. (Eds.), New Perspectives on our Lives with Companion Animals. University of Pennsylvania Press, Philadelphia, Pennsylvania, pp. 47–56.

Hart, L.A., 2000. Methods, standards, guidelines, and considerations in selecting animals for animal-assisted therapy. Part A: understanding animal behavior, species, and temperament as applied to interactions with specific populations. In: Fine, A.H. (Ed.), Handbook on Animal-Assisted Therapy: Theoretical Foundations and Guidelines for Practice. Academic Press, San Diego, CA, pp. 81–97.

Heller, D., Watson, D., Ilies, R., 2004. The role of person versus situation in life satisfaction: a critical examination. Psychol. Bull. 130 (4), 574–600.

Hennessy, M.B., Voith, V.L., Mazzei, S.J., Buttram, 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.

Hoffman, M., 1999. Lend me an Ear: The Temperament, Selection, and Training of the Hearing Ear Dog. Doral Publishing, Wilsonville, OR, 220 pp.

Hsu, Y., Serpell, J.A., 2003. Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. JAVMA 223 (9), 1293–1300.

Humphrey, E.S., 1934. Mental tests for shepherd dogs: an attempted classification and evaluation of the various traits that go to make up ''temperament'' in the German Shepherd Dog. J. Hered. 25 (3), 128–136.

James, W.T., 1951. Social organization among dogs of different temperaments, terriers and beagles, reared together. J. Comp. Physiol. Psychol. 44, 71–77.

John, O.P., Benet-Martinez, V., 2000. Measurement, scale construction, and reliability. In: Reis, H.T., Judd, C.M. (Eds.), Handbook of Research Methods in Social Psychology. Cambridge University Press, Cambridge, England, pp. 339–369.

John, O.P., Gosling, S.D., 2000. Personality Traits. In: Kazdin, A.E. (Ed.), Encyclopedia of Psychology, vol. 6. American Psychological Association, Washington, DC, pp. 140–144.

Keeler, C.E., 1947. Coat color, physique, and temperament; materials for the synthesis of hereditary behavior trends in the lower mammals and man. J. Hered. 38, 271–277.

Ledger, R.A., The Scientific Committee, 2003. Aggressive behaviour in dog breeds re-homed from rescue shelters. In: Ferrante, V. (Ed.), Proceedings of the 37th International Congress of the ISAE, Abano Terme, Italy, Fondazione Iniziative Zooprofilattiche e Zootechnicie, Brescie, Italy, p. 184.

Ledger, R., Baxter, M., 1996. A validated test to assess the temperament of dogs. In: Duncan, I.J.H., Widowski, T.M., Haley, D.B. (Eds.), Proceedings of the 30th International Congress of the ISAE, Guelph, Canada, Col. C.K. Centre for the Study of Animal Welfare, Canada, p. 111.

Ledger, R.A., Baxter, M.R., 1997. The development of a validated test to assess the temperament of dogs in a rescue shelter. In: Mills, D.S., Heath, S.E., Harrington, L.J. (Eds.), Proceedings of the First International Conference on Veterinary Behavioral Medicine, Birmingham, UK, Universities Federation for Animal Welfare, United Kingdom, pp. 87–92.

Ledger, R.A., Baxter, M., McNicholas, J., 1995. Temperament testing dogs in a rescue shelter: improving owner-dog compatibility. In: Rutter, S.M., Rushen, J., Randle, H.D., Eddison, J.C. (Eds.), Proceedings of the 29th International Congress of the ISAE, Exeter, UK, Universities Federation for Animal Welfare, United Kingdom, pp. 101–102.

Lester, D., 1983. Body build and temperament in dogs. Percept. Motor Skill. 56, 590.

Lipsey, M.W., Wilson, D.B., 1996. Practical Meta-Analysis. Sage Publications, Newbury Park, CA, 264 pp.

Mahut, H., 1958. Breed differences in the dog's emotional behaviour. Can. J. Psychol. 12 (1), 35–44.

McCrae, R.R., Costa Jr., P.T., Ostendorf, F., Angleitner, A., Hrebickova, M., Avia, M.D., Sanz, J., Sanchez-Bernardos, M.L., Kusdil, M.E., Woodfield, R., Saunders, P.R., Smith, P.B., 2000. Nature over nurture: temperament, personality, and life span development. J. Pers. Soc. Psychol. 78, 173–186.

Morris, D., 2002. Dogs: The Ultimate Dictionary of over 1000 Breeds. Trafalgar Square Publishing, North Pomfret, VT, 752 pp.

Murphy, J.A., 1995. Assessment of the temperament of potential guide dogs. Anthrozoös 13 (4), 224–228.

Murphy, J., 1998. Describing categories of temperament in potential guide dogs for the blind. Appl. Anim. Behav. Sci. 58, 163–178.

Netto, W.J., Planta, D.J.U., 1997. Behavioural testing for aggression in the domestic dog. Appl. Anim. Behav. Sci. 52, 243–263.

Paulhus, D.L., Bruce, M.N., 1992. The effect of acquaintanceship on the validity of personality impressions: a longitudinal study. J. Pers. Soc. Psychol. 63, 816–824.

Pavlov, I.P., 1906. The scientific investigation of the psychical faculties or processes in the higher animals. Science 24, 613–619.

Pervin, L.A., John, O.P., 1997. Personality: Theory and Research, seventh ed. Wiley, New York, NY, 656 pp.

Podberscek, A.L., Serpell, J.A., 1996. The english cocker spaniel: preliminary findings on aggressive behavior. Appl. Anim. Behav. Sci. 47, 75–89.

Registration Statistics, Retrieved July 26, 2004, from http://www.akc.org/breeds/reg_stats.cfm.

Reuterwall, C., Ryman, N., 1973. An estimate of the magnitude of additive genetic variation of some mental characters in Alsatian dogs. Hereditas 73, 277–284.

Roll, A., Unshelm, J., 1997. Aggressive conflicts amongst dogs and factors affecting them. Appl. Anim. Behav. Sci. 52, 229–242.

Rosenthal, R., 1991. Meta-Analytic Procedures for Social Research. Sage Publications, Newbury Park, CA, 168 pp.

Royce, J.R., 1955. A factorial study of emotionality in the dog. Psychol. Monogr. 69 (22), 1–27.

Ruefenacht, S., Gebhardt-Henrich, S., Miyake, T., Gaillard, C., 2002. A behavior test on German Shepherd dogs: heritability of seven different traits. Appl. Anim. Behav. Sci. 79, 113–132.

Saetre, P., Strandberg, E., Sundgren, P.-E., Petterson, U., Jazin, E., Bergström, T.F., in press. The genetic contribution to canine personality. Genes Brain Behav.

Scott, J.P., Bielfelt, S.W., 1976. Analysis of the puppy testing program. In: Pfaffenberger, C.J., Scott, J.P., Fuller, J.L., Ginsburg, B.E., Bielfelt, S.W. (Eds.), Guide Dogs for the Blind: Their Selection, Development and Training. Elsevier, Amsterdam, pp. 39–75.

Seksel, K., Mazurski, E.J., Taylor, A., 1999. Puppy socialization programs: short and long term behavior effects. Appl. Anim. Behav. Sci. 62, 335–349.

Serpell, J.A., 1983. The personality of the dog and its influence on the pet-owner bond. In: Katcher, A.H., Beck, A.M. (Eds.), New Perspectives on our Lives with Companion Animals. University of Pennsylvania Press, Philadelphia, PA, pp. 57–63.

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.

Slabbert, J.M., Odendaal, J.S.J., 1999. Early prediction of adult police dog efficiency—a longitudinal study. Appl. Anim. Behav. Sci. 64, 269–288.

Stephen, J.M., Ledger, R.A., The Scientific Committee, 2003. Owners are reliable observers of their own dog's behaviour. In: Ferrante, V. (Ed.), Proceedings of the 37th International Congress of the ISAE, Abano Terme, Italy, Fondazione Iniziative Zooprofilattiche e Zootechnicie, Brescie, Italy, p. 190.

Stephen, J.M., Ledger, R.A., Stanton, N., 2001. Comparison of the perceptions of temperament in dogs by different members of the same household. In: Garner, J.P., Mench, J.A., Heekin, S.P. (Eds.), Proceedings of the 35th International Congress of the ISAE, Davis, California, USA, Center For Animal Welfare, UC Davis, p. 113.

Stubbs, C.J., Cook, M., 1999. Personality, anal character, and attitudes toward dogs. Psychol. Rep. 85 (3), 1089–1092.

Svartberg, K., 2002. Shyness–boldness predicts performance in working dogs. Appl. Anim. Behav. Sci. 79, 157–174.

Svartberg, K., Forkman, B., 2002. Personality traits in the domestic dog (*Canis familiaris*). Appl. Anim. Behav. Sci. 79, 133–155.

Topál, J., Miklósi, Á., Csányi, V., Dóka, A., 1998. Attachment behavior in dogs (*Canis familiaris*): a new application of Ainsworth's strange situation test. J. Comp. Psychol. 112, 219–229.

Tortora, D.F., 1983. The Right Dog for You. Simon and Schuster, New York, NY, 381 pp.

van der Borg, J.A.M., Netto, W.J., Planta, D.J.U., 1991. Behavioural testing of dogs in animal shelters to predict problem behaviour. Appl. Anim. Behav. Sci. 32, 237–251.

Wahlgren, K., Lester, D., 2003. The big four: personality in dogs. Psychol. Rep. 92, 828.

Weiss, E., Greenberg, G., 1997. Service dog selection tests: effectiveness for dogs from animal shelters. Appl. Anim. Behav. Sci. 53, 297–308.

Wilcox, B., Walkowicz, C., 1995. The Atlas of Dog Breeds of the World, fifth ed. TFH Publications, Neptune, NJ, 896 pp.

Wilsson, E., Sundgren, P.-E., 1997. The use of a behavior 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.

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.