A

Animal Models

Caio Maximino[1], Saskia S. Arndt[2] and Franz Josef van der Staay[3,4]
[1]Laboratório de Neurociências e Comportamento, Grupo de Pesquisas em Neurofarmacologia e Psicopatologia Experimental, Instituto de Estudos em Saúde e Biológicas, Universidade Federal do Sul e Sudeste do Pará, Marabá, Brazil
[2]Division of Animal Behavior, Department of Animals in Science and Society, Faculty of Veterinary Medicine, University Utrecht, Utrecht, The Netherlands
[3]Behaviour and Welfare Group (formerly Emotion and Cognition Group), Department of Farm Animal Health, Veterinary Faculty, Utrecht University, Utrecht, The Netherlands
[4]UMC Utrecht Brain Center, Utrecht University, Utrecht, The Netherlands

underlying substrates with the final aim to learn about these behaviors and substrates in the species to be modeled. When developing a model or when selecting the most appropriate animal model for a research project, the most important selection criterion is its validity. Validity is the extent to which an animal model's underlying substrates and mechanisms, measurements, and the conclusions drawn from these measurements are well-grounded and represent the real world accurately.

The basic classes of validity, i.e., internal and external validity and face, construct, and predictive validity, will be discussed, plus additional classes and subcriteria of validity that have been proposed as refinement and extension of criteria for evaluating animal models. Also, different types of animal models and a strategy to evaluate animal models systematically are described.

Introduction

In the neurosciences, animal models play a central role for experimentally investigating neurobehavioral (dys)functions and their underlying (patho)physiological mechanisms and processes. Almost always, when using animal models, one implicitly assumes that they simulate these mechanisms and processes in humans or a species other than the one investigated. These animal models thus focus on the homology/analogy of the behavior and

Definition

"An animal model with biological and/or clinical relevance in the behavioral neurosciences is a living organism used to study brain-behavior relations under controlled conditions, with the final goal to gain insight into, and to enable predictions about, these relations in humans and/or a species other than the one studied, or in the same species under conditions different from those under which the study was performed" (van der Staay 2006, pp. 133–134).

J. Vonk, T. K. Shackelford (eds.), *Encyclopedia of Animal Cognition and Behavior*,
https://doi.org/10.1007/978-3-319-47829-6_2059-1

This definition is followed throughout this entry and is basic to the model development and model evaluation process depicted in Fig. 1.

Concepts, Mistaken for, and Concepts that Are Closely Related to "Animal Models"

The animal model literature is filled with conceptual confusions about the concepts test and model. They are categories of animal experimentation that have different purposes, different levels of validity, and different epistemic characteristics (see Table 1). Also, the literature sometimes fails to distinguish between a *model organism* and *an animal model.* A model organism is used to study biological phenomena in a target species, due to specific characteristics that include ease of manipulation and reproduction in the laboratory.

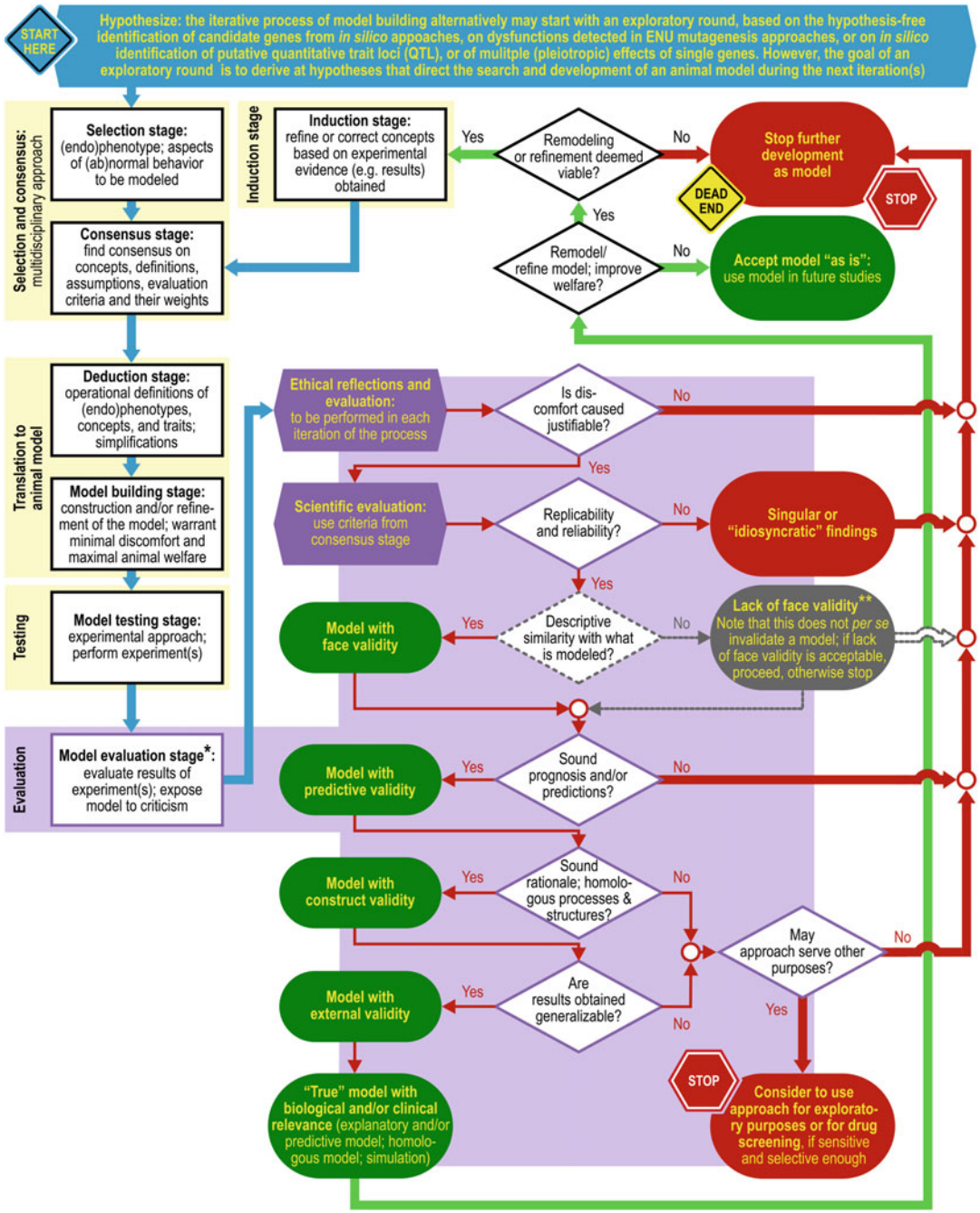
Willner (1991) discriminated three main types of animal experimentation in the behavioral sciences – screening tests, behavioral bioassays, and simulations – that are characterized by different uses. The first two fall under the category of **tests**.

Screening Tests, Behavioral Bioassays, and Simulations

Screening tests are used to predict a desired drug activity from which novel candidates can be assessed. They can be further broken down into two subtypes: "First, some are categorized based on their ability to identify drug candidates with a specific type of clinical action. [...] Conversely, screening tests might also target a particular biochemical action or property in order to identify new drug candidates" (Wright 2002, p. 135). Screening tests have two main requirements: that they make exact and effective predictions about the biochemical action of the drug candidates and that they show that the candidates are efficacious in the target species or disorder (Willner 1991). Screening tests require only predictive validity, and the "relevance of the procedure [. . .] is not an issue, as long as the goal of correctly identifying drugs and drug effects is achieved" (Wright 2002, p. 135). The main use of a screening test is in pharmacological and pharmaceutical research.

Behavioral bioassays are used to study the physiological and neurobiological mechanisms that are associated with brain function, being associated with either the evaluation of specific physiological systems or neuronal circuits or measuring dysfunctions produced by experimental manipulations (Willner 1991). Given that behavioral bioassays do not simply seek to develop novel drugs, these types of animal experimentation can have increased explanatory power over screening tests.

Simulations generally attempt to model mental disorders based on comparative studies of the same states and conditions (Willner 1991). Simulations are more commonly used to study the pathophysiology and treatment of mental disorders, but they are also useful to produce insights into normal function. Simulations should require the use of animals with pathological organisms and therefore are highly dependent on the criteria and restrictions of validity (see below). *Only a simulation is a proper model, and both screening and behavioral bioassays should be considered "*tests.*"* While some behavioral bioassays use pathological organisms (e.g., the olfactory bulbectomy test, in which ablation of the olfactory bulb produces learning deficits, hyper-reactivity, and glucocorticoid responses which are reversed by antidepressants or the stress-induced hyperthermia, in which transferring rats or mice to a novel environment increases body temperature in an anxiolytic drug-sensitive way), it is not *required* of them that the induction methods are analogs of etiological factors of the target disorder. Similar observations can be made for most research on transgenic organisms (e.g., knockout rodents). *Across this entry, "animal model" will be used interchangeably with "simulation."* Table 2 provides a non-exhaustive overview over different types of model animals and of the dependent and independent variables in animal models.



Animal Models, Fig. 1 The iterative process of model building and model evaluation. (Figure modified from van der Staay et al. 2014 with permission). ∗: Here, face, predictive, and construct validities are used as criteria to evaluate a model. However, one may use other sets of validities instead. For examples of additional/alternative types of validity that could be used to evaluate a model (see e.g., Belzung and Lemoine 2011, Ferreira et al. 2019, and Hoffman 2016). ∗∗: If face validity is considered as crucial, failing to meet this criterion should end the model building and evaluation process

Animal Models, Table 1 A classification of behavioral tests and models based on purpose, epistemic characteristics, and main use

Category	Type, based on Willner (1991)	Purpose	Epistemic characteristics	Main use
Tests	Screening tests	Allows limited comprehension of an isolated aspect of a complex biological mechanism to be mapped Limited predictive usefulness (e.g., predicting desired drug activity)	Underlying biological mechanism does not need to be similar Not necessarily sensitive to, e.g., triggering factors Not necessarily hypothesis-driven Low construct validity; moderate predictive validity (pharmacological isomorphism only)	Drug discovery and screening
	Behavioral bioassays	Allows broader comprehension of a mechanism, without necessary causal analogy Moderate predictive usefulness [e.g., studying neural bases of behavioral (dys)functions]	Biological mechanism is similar, but not causally analogous By definition, is sensitive to disturbances High predictive validity, low to moderate construct validity	Investigation of the physiological mechanisms that compose and organize brain function
Models	Simulations	Can allow inferences and extrapolation to the human disorder, with high probability that the hypothesis thus generated is true	Similar mechanisms with probable causal analogy By definition, sensitive to disturbances High face, predictive, and construct validity (considers the need to address theoretical constructs on the etiology, symptomatology, and treatment)	Investigation of the neurobiological underpinnings of a disease

Biological and Clinical Relevance

An important criterion for using a behavioral model is biological and/or clinical relevance (van der Staay 2006). Relevance can be determined using many different criteria (van der Staay 2006), all of which depend on the purpose of the model, the information available for the evaluation, and the stage of development of the model (see Fig. 1). Thus, tests and models have different uses, which should be explicitly defined at the outset: "The explicit definition and designation of the specific purposes that an animal model should fulfill is basic as it allows to define a set of weighted criteria for evaluating the model" (van der Staay et al. 2009, p. 2). Many different uses of a model, and the differentiation between tests (screening, bioassays) and models (simulations), suggest that "(. . .) the value of an animal model is relative, not necessarily to the degree that it simulates the structural, functional, and behavioral correlates of [psychopathology], but to the degree that it comports with the specific tasks for which it is being used. The aims of neuropsychopharmacological research are heterogeneous, often being contextually calibrated to fit specific needs and questions of a wide variety of researchers. Consequently, acknowledging the proper roles and capacities of animal models, and not overgeneralizing or wresting them from their intended domains of application, is therefore a major step in arresting criticism pertaining to the legitimacy and logistics of animal modeling" (Wright 2002, p. 134).

Animal Models, Table 2 Types of model animals and the independent and dependent variables in animal models. (Modified from van der Staay et al. 2009, p. 4. Table 1A, B). Part A lists different types of "model animals" for the study of behavioral dysfunctions (inspired by Gamzu 1985). It focuses on the type of subject (independent variable) and is not concerned with the type of dependent variable measured

A: Types of "model animals" (note that the list of models is not exhaustive)

"Normal" animals, i.e., animals without any observable behavioral deficit. (Studying behavior in "normal" animals provides the baseline data for identifying abnormal behavior)	**Animals with naturally occurring neuronal deviations, neuropathologies, or neurological/psychiatric (endo)phenotype(s)**	**Animals with experimentally induced behavioral dysfunctions or characteristic behavioral (endo)phenotype(s)**
Note: At first glance, normal animals are not model animals for the study of, for example, neurobehavioral disorders. Normal animals are used to: • Assess the safety/toxicology risk of a putative therapeutic; • Obtain an estimate of the putative abuse liability of a compound • Explore the neurobiological specificity of compounds and their molecular and cellular mechanisms of action Isogenic populations, for example: • Monozygotic twins or multiples • Clones • Inbred strains and variants; coisogenic strains (differing at one single locus); Coisogenic strains (differing by a mutation at a single locus); congenic strains, (differing at only one locus and a linked chromosomal segment); recombinant inbred strains (with a permanent set of recombinations between chromosomes from two or more inbred strains) Non-isogenic populations, for example: • Random bred and outbred strains; animals caught in the wild	Spontaneously and endogenously occurring psychiatric or neurological conditions; spontaneously occurring mutations; aging animals Genetic lines: Inbred strains and their crossings Selection lines, established through artificial selection favoring low and/or high values of a particular trait within a given population for a number of successive generations; preferentially including an unselected control line for comparison Selected extremes from a particular animal population, e.g., good vs. poor learners, dominant vs. subordinate animals, non-aggressive vs. aggressive animals; anxious vs. bold animals	Transgenic and knockout animals; isogenic or non-isogenic populations that express a neurobehavioral deficit or a neurological/psychiatric (endo)phenotype Animals from mutagenesis screens (after thorough phenotyping and validation) Selection lines resulting from selective breeding Environmental factors: e.g., animals experiencing acute or chronic stress, pain, or sleep deprivation Animals with disruptions induced by: • Dietary composition (e.g., tryptophan depletion) • Pharmacological compounds (e.g., scopolamine, MK-801), or • Electrically, or by hypoxia, or anoxia Animals with focal or global ischemic, embolic, or hemorrhagic cerebral stroke Animals with CNS-specific lesions, for example: neuro- or immunotoxic lesions; lesions induced by aspiration, ablation (knife cuts); radio-frequency lesions, cryogenic lesions

B: Independent and dependent variables

Independent variable (in essence the model animals and their controls listed in part A)	**Dependent variables and/or (endo)phenotypes**	
	Neuropathological changes	**Behavioral changes**
E.g., genetically modified animals, aged animals, lesioned animals, ischemic animals, hypoxic animals, aged and lesioned animal, i.e., combination of deficits (see part A)	Damage or dysfunctions induced: Site and size of neuronal damage (neuropathology), effects on specific neuronal circuits or neurotransmitter systems, psychophysiological and/or biological (endo)phenotypes	Behavioral dysfunction or malfunction: Impaired cognitive performance, impaired sensorimotor functions, neuropsychiatric symptoms, behavioral (endo)phenotypes
	Homology of damaged area(s) or neuropathological changes.	Homology of disrupted processes or impaired functions
	Operational definition(s) of the neuropathological (endo)phenotype(s)	Operational definition(s) of the behavioral (endo)phenotype(s)

An important consequence of that formulation is that "relevance" is always relative to the aims of the research and not an a priori property of the model. "Relevance reflects the meaningfulness and usefulness of results obtained with an animal model for a particular scientific and/or clinical purpose" (van der Staay 2006, p. 146). The purposes of a model are to enhance our understanding of the underlying substrates and mechanisms associated with the disorder and to assess the effects of drugs and treatments in a different system than the target population (e.g., the patient) (Maximino and van der Staay 2019; van der Staay 2006), allowing inferences and extrapolation to disorder in a certain target species, with high probability that the hypothesis that is generated by the model is true (Maximino and van der Staay 2019). However, a model may have a more limited applicability but still may be relevant for specific purposes (van der Staay et al. 2009).

While the concept of relevance is central to this formulation of behavioral models, it is sometimes confounded with the concepts of validity (see below). Thus, models possess relevance if they are valid and generalizable (van der Staay 2006). Since behavioral bioassays and screening tests do not require construct validity, their relevance lies in their sensitivity to find novel therapeutics (in vitro and in vivo) and/or in their capacity to generate novel hypotheses on normal and pathological brain functioning.

Model Building as an Iterative Process

The aspects discussed so far draw attention to an implicit aspect of the use of animal models, viz., that model building is an iterative process. Therefore, it is sometimes difficult to judge whether a model is established (i.e., that it satisfies the preset validation criteria) or whether its developmental process is reaching a dead end because it fails to meet essential validation criteria (see Fig. 1). The central role of theory in construct validity (Maximino and van der Staay 2019), as well as the importance of the *zeitgeist* in the creation of a model, creates the necessity to make the theoretical basis of the model explicit. All model development begins with a hypothesis derived from data or theories (van der Staay 2006). Model building proceeds from this hypothesis-building stage to a **selection stage**, in which aspects of normal or abnormal behavior of the species to be modeled are selected. "This stage thus reflects the intended animal model and its first translation into testable behavior that is believed to be indicative for the behavioral (dys)function to be modeled." (van der Staay 2006, p. 142). The next step in the model building process is a **consensus stage**, in which the criteria, definitions, and assumptions regarding the validity of the analogy are discussed and agreed upon. The following stage, the **deduction stage**, involves the decomposition of the endophenotypes under analysis and the verification of the translational relevance (i.e., whether the same endophenotype can be tested in both the model and the target) (Fig. 1). These steps are mostly theoretical and are followed by the **model building stage** per se (van der Staay 2006).

Probably the most important stage in this process is the **model evaluation stage** (Fig. 1), in which the results obtained with the model are evaluated (van der Staay et al. 2009). The many different proposals for evaluation strategies will be discussed in the next section. If the model is considered acceptable after this stage, the process enters an **induction stage**, in which knowledge gained from the experiments performed to develop and validate the model can be used to further refine or correct the concepts which sustain the model; otherwise, the researchers should determine whether the failure was due to choosing inappropriate criteria to build the model or whether the criteria need to be refined.

Note that not all criteria can be evaluated in a single experiment. Testing replicability, for example, needs data from at least two experiments. For the time being, therefore, one may assume that a criterion is fulfilled, until enough evidence has been accumulated (i.e., after a number of experiments have been performed) to base the decision about whether a particular criterion is met, or whether the model fails to meet a criterion, on scientific evidence.

Model Evaluation and Concepts of Validity

Basic Classes of Validity

Concepts of validity are highly variable between model theorists (Belzung and Lemoine 2011; Ferreira et al. 2019; van der Staay 2006). While most follow the recommendation of using face, construct, and predictive validity as the main or essential criteria for evaluating a model, many other propositions appeared; Table 3 attempts to summarize these suggestions, which are detailed in what follows.

Internal Validity and Reliability

Among the proposed methods for evaluating a model is the assessment of its reliability and replicability. **Reliability** is primarily a property of the method, while **replicability** is a property of the results obtained using a particular model. The first, then, indicates how consistent the method is or the extent to which the model yields consistent results each time that measurements are performed under the same conditions. In animal modeling, reliability refers to at least three different properties: the replicability of the model in different laboratories; the concordance of measures taken by different independent observers or by different methods (e.g., automated vs. manual data collection); and the evaluation of the temporal stability of the behavior measured. The concept is conflated with replicability, "the degree of accordance between the results of the same experiment performed independently in the same or different laboratories" (van der Staay et al. 2009, p. 3). Indeed, test theory identifies several classes of reliability estimates (inter-rater reliability, test-retest reliability, inter-method reliability, internal consistency reliability) which are sometimes related to replicability.

Both reliability and replicability are part of the **internal validity** of an experimental method (Table 3; Fig. 2), a characteristic of "how well a study was performed, how strictly putative confounding variables were controlled, and how confident one can be that the changes observed in the dependent variable(s) are *caused* by experimentally manipulating the independent variable(s), and not by confounds" (van der Staay et al. 2009, p. 3). Internal validity refers mainly to the need to objectively, reliably, and replicably measure the phenomena being evaluated; therefore, it refers to the consistency of the experimental design (Belzung and Lemoine 2011) and the reduction of biases through randomization and blinding (Ferreira et al. 2019) as sine qua non conditions for external validity. Strictly speaking, internal validity is also about causality (van der Staay et al. 2009).

Face Validity

Face validity refers to whether the test has a perceived resemblance to its target – whether it "looks like" its target. One important limitation to the attempt to establish face validity as a point-to-point correspondence between a disorder and an animal model is "the fact that there is no good reason to suppose that a given condition will manifest itself in identical ways in different species" (Maximino et al. 2010, p. 117). The presence of a symptom in itself has no value if it is not assumed to indicate the expression of an underlying pathological condition, and many authors consider face validity of lower significance for the evaluation of a model (McKinney and Bunney 1969; van der Staay et al. 2009).

Discarding direct references to symptomatology and focusing on neurobehavioral alterations that are distinctive of a given disorder increases face validity because it allows for the use of a given *endophenotype* that is observed in both species. For example, schizophrenic patients have deficits in working memory, attention, and sensorimotor gating that are not part of the diagnostic criteria; since sensorimotor gating can be assessed via measures of prepulse inhibition (PPI) of the startle reflex, decreases in PPI in nonhuman animals are often used as evidence for face validity (Gould and Gottesman 2006). This strategy shifts attention toward other variables, but the functional significance or the specificity of the endophenotype is not a requirement for its definition. For example, not all schizophrenic patients show PPI deficits, and PPI alterations can be found in other neuropsychiatric diseases.

Face validity is also increased by focusing on cross-species psychological processes, a strategy

Animal Models, Table 3 Concepts of validity, their definition, and their relations. (Adapted from Belzung and Lemoine 2011, Maximino et al. 2010, McOmish et al. 2014, and van der Staay et al. 2009)

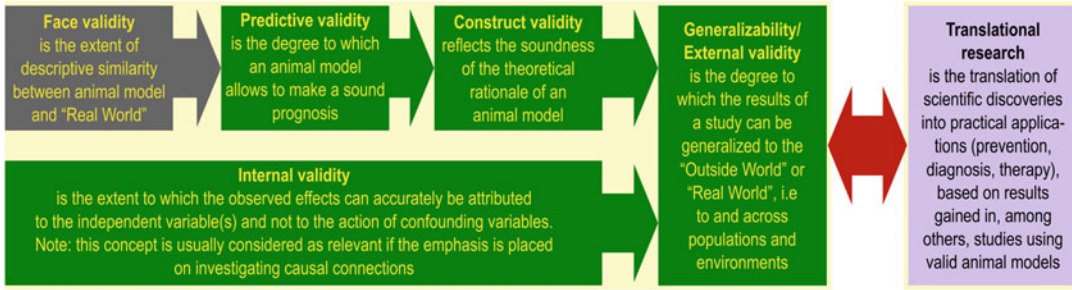
Validity concept		Definition
Internal validity	Replicability	"The degree of accordance between the results of the same experiment performed independently in the same or different laboratories" (van der Staay et al. 2009, p. 3)
	Reliability	The overall consistency of measurements made using the model
Face validity	Ethological validity	Similarity of behaviors in normal and pathological organisms (Belzung and Lemoine 2011)
	Biomarker validity	Also an aspect of convergent validity; whether biomarkers associated with the disease also appear in the model (Belzung and Lemoine 2011)
Predictive validity	Pharmacological isomorphism/remission validity	A given drug effect being characteristic for a well-defined class of chemicals and indicative of a specific mode of action
	Induction validity	"The action of the etiological factors on the observable effects of the model disease resemble its action on the observable effects of the human disease" (Belzung and Lemoine 2011, p. 9)
Construct validity	Homological validity	Whether the trait in question is homologous in humans and the species tested (Maximino et al. 2010)
	Genetic validity	Similarity of the genetic or genomic basis for the endophenotype / disordered domain
	Environmental validity	Similarity of G x E interactions (McOmish et al. 2014) Ontopathogenic validity: similarity of "early environmental factors whose interaction with the initial organism produces a vulnerable organism" (Belzung and Lemoine 2011, p. 7); triggering validity: "the similarity of triggering factors occurring during adulthood whose interactions with either a vulnerable or an initial organism produces a pathological organism" (Belzung and Lemoine 2011, p. 7)
	Mechanistic validity	Similarity of the mechanism one supposes or knows is working in the animal disease to the mechanisms that is presumed to be working in the human disease
	Pathogenic validity	Similarity of the processes that lead to disease
	Inter-relational validity	Ability of the proposed model to target the interplay between various disordered domains (Stewart and Kalueff 2015)
External validity	Convergent validity	Similarity of results from different methods
	Translational value	Ability of the proposed model to produce results which are translated to the target disorder

that "(. . .) requires prior extensive research on the biology and ethology of these species" (Belzung 2014, p. 1044), and is referred to as **ethological validity**. Ethologically valid models use endpoints that are functionally meaningful, serving the same *function* as the target behavior, domain, or endophenotype. Perhaps no field has focused so much on this as the field of anxiety research (Maximino et al. 2010).

Belzung and Lemoine (2011) proposed that face validity is associated with both ethological validity and **biomarker validity**: "The former is the similarity of behaviors related to the presumed pathological organism. Here again, the meaning matters more than the material similarity. Biomarker validity is the similarity of biological markers related to the presumed organism. What matters is the function of a marker, not its chemical composition" (Belzung and Lemoine 2011, p. 9).

Predictive Validity

Predictive validity relates to the success of predictions derived from the model. Willner (1991) emphasized that predictive validity also includes McKinney's and Bunney's (McKinney and Bunney 1969) "similarity in treatment effectiveness"; in that



Animal Models, Fig. 2 Face, predictive, construct, and external validities are the basis for animal model development and model evaluation (see also Fig. 1). In addition, in experimental approaches, two classes of validity may be distinguished: internal and external validity. The bidirectional and recursive relationship between animal models, translation to applications, and reverse translation to animal models is indicated by the two-headed arrow to translational research. Insights gained in the translational stage may also feedback (reverse translation) to the model building and model evaluation process

sense, predictive validity reduces to *remission validity* or *pharmacological isomorphism*. While theoreticians for model building agree that predictive validity is less important than construct validity (van der Staay 2006; van der Staay et al. 2009), in practice most researchers appear to consider pharmacological isomorphism as more important (Maximino et al. 2010). Pharmacological isomorphism also requires that drugs which are *not* effective in clinical settings do not affect the model (i.e., no errors of commission are present); for example, amphetamines are commonly used in pharmacological validation of anxiety models because they do not reduce anxiety, but can produce "false positives" by altering locomotion (Willner 1991). The appeal to pharmacological isomorphism, usually without reference to construct validity, weakens the operationality of the model and risks circularity in the definition; for example, defining an outcome in the EPM as anxiety-like because anxiolytic drugs decrease it, although a common practice, is circular definition.

Pharmacological isomorphism can be an important criterion for screening tests and behavioral bioassays, but it is not sufficient for a simulation (Willner 1991).

In addition to showing responses to clinically effective drugs, predictive validity of proper models requires that the model is responsive to drugs and treatments that exacerbate the same symptoms and that the model is responsive to all classes of drugs that are useful in treating the target disorder instead of being only sensitive to the effects of drugs with specific pharmacological characteristics (e.g., structure-activity studies) or show only narrow effects on specific neuronal functions (e.g., serotonin reuptake inhibitors). Also, predictive animal models should be sensitive to the relative potencies of different drugs corresponding with their potencies in clinical use (Ferreira et al. 2019). Moreover, simulations are less sensitive to false positives and false negatives than screening tests and behavioral bioassays: "False positives may be clinically effective at high doses, but go unnoticed due to the desire to stave off dire side effects. On the other hand, false negatives raise the possibility of clinical subgroups and populations with differing response patterns" (Wright 2002, p. 137).

Construct Validity

Construct validity refers to the theoretical rationale behind a given model. The most influential description of construct validity in animal modeling is that of Willner (1986), who proposed five subcriteria for a model to have construct validity: "[w]hether both the behavior in the model (1) and the features of [the disorder] being modeled (2) can be unambiguously interpreted, and are homologous (3), and whether the feature being modeled stands in an established empirical (4) and theoretical (5) relationship to [the target disorder]" (Willner 1984, p. 1).

Subcriteria for Construct Validity

There is no consensus about what exactly Willner's (1984) sub-criterion 5 entails. Atanasova (2015) suggests that construct validity "requires proper mapping between the hypothesized mechanisms underlying the human and the animal behaviors respectively" (Atanasova 2015, p. 172). Belzung and Lemoine (2011) argued that the theoretical relationship between model and target is understood in a polysemic sense, including theories about the nature of the disorder; the centrality of a dysfunctional process to the disorder (e.g., helplessness and anhedonia are central symptoms of depression, and fear generalization is a central aspect of PTSD, etc.); the dynamic of the disorder (time course, relationship to previous stress, etc.); and etiological aspects. Etiological aspects, in turn, include theories on the role of triggering events in causing the disorder in human patients and in animal models (e.g., stress may cause depressive-like symptoms in animal models and major depressive disorder in humans); the centrality of specific characteristics of these triggering events (e.g., stressor uncontrollability, intensity, or distance to threat); and the involvement of associated biological processes (e.g., dysfunctions in the brain reward system or the fear system).

Another important aspect of the theoretical grounding of construct validity is that properly developed theories can provide the model with stronger predictive and face validity (McNaughton and Zangrossi 2008). As a consequence, novel models generated directly from a theory can provide "a range of constructs from which a suitable model can be extracted by design" (McNaughton and Zangrossi 2008, p. 17).

The influence of the *zeitgeist* is especially important when considering the theoretical basis of construct validity. It has been argued that the theoretical choices that constrain model building should not be understood as homological – that is, "the choice for a model based on operant or respondent behavior, in a naturalistic situation or related to pharmacologic (sic) manipulation," should not imply that the pathology that is modeled is reducible to these mechanisms (Gouveia and de Brito 2015, p. 308). Nonetheless, theoretical considerations are very important in model building: "Model-based science is an exercise in *theorizing* or speculating on unknown mechanisms and outcomes. Modeling is always an indirect type of theory-building practice" (Degeling and Johnson 2013, p. 95). A properly developed theory encapsulates a network of data and other theories, and by capturing the "essence" of this network, the construct is validated (McNaughton and Zangrossi 2008).

One central theoretical assumption of animal models is that of homology between model and target, in the evolutionary sense (Greek and Rice 2012). The (usually unspoken) assumption is that, since mental (and nonmental) disorders represent an extreme in a continuum, the trait/phenotype/domain of interest is conserved between healthy humans and other species. For example, it has been proposed that "normal" anxiety is an adaptive response that is part of the defensive repertoire of the animal. The obvious consequence is that the normal defensive repertoire of a given animal can be used to study "normal," non-pathological anxiety (or at least some aspects of anxiety), as well as normal fear-like behavior. That is the concept of **ethological validity** proposed by Belzung and Lemoine (2011). Usually, however, the conservation is implied but never tested (Maximino et al. 2010). Moreover, which endpoint to choose cannot be so easily determined for other domains that are dysfunctional in several disorders, such as cognition.

External Validity, Standardization, and Generalization of Tests and Models

In addition to the earlier described well-accepted criteria, other researchers have suggested the application of additional validity criteria to improve the ability of animal models to identify novel drug targets (which is not the same as enhancing the analogy). Thus, genetic validity, environmental validity, and population validity have been proposed as important criteria (Hoffman 2016). These are special cases of **external validity**, which "involves an inference to the *robustness* of a causal relation outside the narrow circumstances in which it was observed and

established in the first instance" (Guala 2003, pp 1198–1199, emphasis in the original). In that sense, external validity is a consequence of construct validity.

The establishment of the external validity of a given model is not straightforward; evaluation of generalizability is usually understood as the last step in the model evaluation stage (Fig. 1). While epistemological discussions on external validity usually rely on subject and laboratory conditions (Guala 2003), in the case of animal models, the tests for external validity rely on whether the model possesses validity across different housing conditions and laboratories or across different behavioral tests that are believed to measure the same underlying traits and whether the model enables insights and predictions to other species (van der Staay et al. 2009).

Genetic and Environmental Validity

Genetic validity, as proposed by Kas and colleagues (2007), refers to whether there is an identical genetic basis for the endophenotype – that is, whether the genes that determine behavior in the model are also associated with the target disorder and shared between humans and the model organism. The use of behavioral domains (endpoints or endophenotypes representative of naturally occurring behavioral functions, such as social interaction, appetitive motivation, memory function, and emotionality) is similar to the "research domain criteria" (RDoC) approach proposed by the NIMH (Cuthbert 2014). In this case, the focus is not on altered (dysfunctional) endophenotypes, but instead on naturally occurring behavioral domains that are relevant to both animal behavior and human neuropsychiatric disorders. It is expected that "inter-species trait genetics rather than complex syndrome genetics will optimize genotype-phenotype relationships for neuropsychiatric disorders and generate powerful new disease models that will facilitate the identification of biological substrates, their developmental consequences, and the environmental modulators underlying these disorders" (Kas et al. 2007, p. 324).

To increase the genetic validity, it is important to model a disordered domain across numerous species (Kas et al. 2007), largely based on the notion that behaviors and their genetic underpinnings are evolutionarily conserved across different species because of common adaptive mechanisms (Greek and Rice 2012). An important limitation of this approach is that genetic validity cannot be properly established for a wide variety of endophenotypes and genes. For example, some genes and behaviors do not correlate across species (LaPorte et al. 2008) – especially considering nonmammalian species, which show different number of genes encoding specific proteins. Moreover, the individual domains and endophenotypes which are analyzed may only incompletely model the target disorder, and in this case, genetic validity can only be established as an accessory argument after construct validity has been established. Finally, genetic analyses in animals usually demand isogenic lineages based on inbred strains, but the human population is genetically heterogeneous, and therefore using only inbred animals decreases external validity (see below) (Stewart et al. 2015).

Environmental validity adds a novel layer to the complexity of genetic validity in requiring that not only gene variants can be identified in both the model and the target disorder but that specific gene by environment interactions can be found for both cases (McOmish et al. 2014). In particular, behavioral models could benefit from "a complementary 'enviromics' approach, in which the 'envirome' contributing to the pathogenesis and progression of each disorder, and multiple disorder ('environmental pleiotropy'), is systematically catalogued and computationally analysed" (McOmish et al. 2014, p. 4723).

The same critique made for the role of etiology in establishing validity can be made for both genetic and environmental validity, though, usually, it is precisely the genetic, genomic, or "enviromic" underpinning of the disorder that is under study with the model, and therefore it is not known beforehand. A solution to this conundrum is assuming that both genetic and environmental validity are "accessories" in establishing external validity and are in fact practical consequences of using different species to test hypotheses. Indeed, the idea that gene by environment relations are

important to the establishment of psychiatric disorders has a rich theoretical background that must be taken into consideration when assessing a model's validity (van der Staay 2006).

Another approach to environmental validity is to consider whether standardized housing appropriately models the conditions in which target species (humans and nonhuman animals) usually live. What usually is referred to as "standard housing conditions for rodents" is in fact barren environments that produce animals that are metabolically morbid, and the environments in which animals are reared lack sensory, cognitive, and motor stimulation (Burrows et al. 2011). This, of course, contrasts with the majority of humans who receive a wide range of sensory, cognitive, and motor stimulation throughout their lifetimes: "The implicit argument here is that the conditions under which laboratory rodents are kept are well below what any wild animal in its natural environment might experience and, having observed the dramatic effects of marginally enriched environments upon the brain chemistry of these animals, we cannot ignore the possibility that a brain which enjoys even some basal level of stimulation (such as a mouse with even minimal levels of enrichment and/or the majority of humans, and thus the target population we are attempting to model) may be predisposed to respond differently to the therapeutics which are being tested preclinically than mice housed under standard conditions" (Burrows et al. 2011, p. 1380).

Important experimental confounds can be thus introduced, reducing the internal validity of the model; moreover, a model that is applied to animals reared in impoverished environments lacks construct validity (Burrows et al. 2011). While concerns on behavioral variability generated by environmental enrichment, as well as increases in cost and space needs, are common difficulties in most research settings, it has been proposed (Burrows et al. 2011; McOmish et al. 2014) that a secondary screen is added to the set of experiments in order to test whether a given intervention that was shown to have efficacy under barren housing conditions also has effects in animals housed under enriched environments.

Convergent Validity

One important aspect of external validity, in the case of animal models, is **convergent validity** (Atanasova 2015). This refers to the ability of results to converge (or at least be compatible) with results obtained with different methods that measure similar traits. Most guidelines for reporting or conducting in vivo animal research make recommendations regarding convergent validity, insisting on replication in different models of the same disease (e.g., different transgenics, strains, or lesion techniques); independent replication by different researchers or, ideally, research groups; and replication on different species. For example, in behavioral genetics it is common to demonstrate that knocking out a given gene produces alterations in more than one behavioral model. In behavioral models with several dependent variables, convergent validity is harder to establish, since the variables do not necessarily measure the same underlying construct (see above). In that case, a variety of multivariate statistical methods can be used to establish the underlying factors/traits that are shared by different dependent variables across tests. Factor-analytical, hierarchical clustering, multitrait-multimethod tests, animal model arrays, and multidimensional scales are useful in that direction (Maximino et al. 2014).

Perhaps more important than convergent validity is the need to increase translational value, the ability of a given model to produce results that are translated to the target condition/disorder. While drug development is not the only aim of animal modeling, the success rates for drugs which move in the production pipeline from preclinical to clinical research remain low, despite important advances in the use of high-throughput approaches to target identification. Belzung (2014) identified four "drawbacks" in current model research that decrease translational value: lack of homological validity; lack of pathogenic validity; lack of refinement of preclinical situations; and predictions from animal models not taken into account when translating findings to the target disorder. In addition to these limitations, the inability to translate has been associated with differences in treatment modalities, when delays

to start treatments that are unrealistic in the clinic are applied in the model; when the dose range used in the model is unfeasible in the clinic due to toxicity or reduced tolerability; or when differences in the timing of testing exist between animal studies and clinical trials (e.g., by using acute versus chronic treatment). Important factors for pharmacological research include treatment initiation time, duration, and frequency, as well as endpoint evaluation time (Stewart et al. 2015). The lack of pathogenic validity, for example, can occur due to the induction of the disease in animals that are healthy when the disorder occurs mainly as a comorbidity; to the use of homogeneous groups of animals, when the disorder affects a heterogeneous group of patients; or to the use of male animals when the disorder occurs mostly in female patients. Most of these recommendations rely on knowing aspects of the disorder that are relevant to the model and incorporate epidemiological data on the model. Thus, the external validity of a model will largely be determined by disorder-specific factors.

External Validity and Standardization

Recently, the low external validity of animal models, particularly in pharmacological research, has been thoroughly discussed, and solutions to increase it have been proposed. The need to standardize procedures has also recently extensively been discussed. One criticism was that task variations sometimes are introduced to study specific aspects of the disordered domain, and therefore using standardized tasks can preclude these specific uses. Stewart et al. (2015) made methodological recommendations for anxiety research that can be applied to a wider field. They recommend focusing on multiple endophenotypes or endpoints, expanding beyond the usual breadth of tests which are common for a disorder. Stewart and co-workers (Stewart et al. 2015) also recommend including a panel of outbred strains to increase "population validity," considering the genetic variability of human populations. They also recommend using "smart" test batteries and considering a wider spectrum of drug regimens beyond the "classical" acute and chronic treatments, including the addition of intermittent treatment, subchronic treatment, and early or late treatment initiation times and the introduction of delayed response evaluations (e.g., assessing results several weeks after treatment initiation). Finally, they recommend clarifying and rethinking "inclusion/exclusion criteria" by including anomalous data (e.g., outliers) in the analysis or considering the potential stratification of these data on the population's responses to drug treatments.

Denayer et al. (2014) proposed introducing "clinical trial features" in animal research. They suggest, in consonance with Stewart et al. (2015), that assessing and reporting "negative" effects (pain, discomfort, dyskinesia, and other side effects) in preclinical research can improve the rate with which potential candidates are discarded before they progress in the pipeline. They also advocate adding predictive biomarkers for treatment responsiveness, such as pharmacogenomic data or peripheral biomarkers, as classifiers in the screening. Finally, they suggest that "reverse translation" is as important as other strategies to refine the model, as biomarker and endophenotype research in patients can inform further iterations of the model building and evaluation processes. Notice that these recommendations point to a feedback between external validity and predictive validity.

A final recommendation regards not the alteration or refinement of animal models but the addition of "nontraditional" stages in clinical trials by testing drugs in an experimental setting in patients. For example, startle reactivity, carbon dioxide sensitivity, and fear generalization have been shown to be altered in anxiety and trauma and stress-related disorders and can be used, in addition to classical tests, to study drug and/or gene effects in rodents (Hoffman 2016). Importantly, they can be used in human patients to test drug effects (Hoffman 2016). Thus, drug candidates that normalize the activity of these "back-translated" variables in animal models would be tested in human subjects using analogous test paradigms in addition to the clinical settings.

Hypothesis-Free Model Building?

The focus on construct validity, with a strong theoretical basis, represents an important epistemological contribution, but does not necessarily cover everything that the scientists that use models and tests do. High-throughput techniques in neuroscience and biomedical research have produced a plethora of genomic and proteomic data which were not driven by a specific hypothesis and are rather exploratory in nature. These "Big Data" approaches include genome-wide association studies in animals and humans, microarray and RNA-Seq studies in animals, and other next-generation sequencing techniques. These strategies analyze thousands of genes or transcripts at the same time, in a "shotgun" approach to attempt to find polymorphisms that are associated with a disorder or genes that are up- or down-regulated after a given treatment or experimental manipulation. As a result, many potential targets which were not predicted by theoretical accounts appear.

In spite of the excitation that these studies generate, the pathophysiology of mental disorders is unlikely to be related to a mutation in one single gene, but rather to the effects in complex biological pathways which involve multiple interacting gene products (Kas et al. 2011). As a result, mechanistic and translational approaches to link the hits obtained using these high-throughput genomic/transcriptomic approaches and genome-wide association studies with relevant endpoints are necessary, even when the candidates were generated with these hypotheses-free approaches. Tactical approaches include the integration of multiple datasets from human and animal studies to identify converging evidence for the role of a given gene or pathway in that disorder (Kas et al. 2011).

While high-throughput genomic/transcriptomic approaches and genome-wide association strategies can certainly increase throughput, they are not without problems. While genetic validity is starting to figure as an important aspect of model building, using it as a starting point to develop a model does not preclude other aspects – including knowledge about species differences and context-specificity of behavioral strategies that are used in a given test (Kas et al. 2009). While using an endophenotype can increase external and face validity because these units are more easily translatable between humans and non-human animals, the choice of a behavioral endophenotype is still subject to other aspects of construct validity, especially homological validity, as well as aspects of face and predictive validity. For example, it is not clear whether the elevated plus-maze is an adequate test for panic attacks, although it can more properly model the hyperanxious aspect of panic disorder. In that sense, prioritizing genetic validity as a starting point is a useful strategy but should be approached carefully in order not to fetishize the gene and forget other aspects of behavioral modeling.

One important limitation of these approaches is that starting from the gene(s) instead of the disorder often creates the need for exploratory research on a plethora of models. This difficulty has already been established in relation to the needs imposed by phenotyping research in genetically engineered mice: "This vast murine factory line (. . .) is demanding ever-faster analysis. And the answers are sought now, ideally in the form of easily set-up, easy-to-follow, standardized protocols with high test-validity and secure intra- and inter-laboratory reliability" (Crabbe and Morris 2004, p. 1175). The authors exemplify, with the water maze, a task that is used to assess spatial memory. Pressures toward standardization and automation of data collection exist and are justified by the need to increase throughput; however, variations of the task allow for the analysis of different aspects of spatial learning and memory, and some of these variations are labor-intensive and preclude a high number of replications. Again, whether a standardized, automated, high-throughput version of the task is ideal depends on its use.

This use of automated tests is sometimes lauded as an advantage; however, besides risking missing important variables, "the suggestion that automation is *per se* a good thing seems to be driven by the belief that automatically recorded data escape the influence of the experimenter. There are scarce data to support this idea"

(Crabbe and Morris 2004, p. 1177). The important factor here should not be throughput, but the appropriate interpretation of behavioral data "that is crucial to their validity" (Crabbe and Morris 2004, p. 1177).

In addition, "hypothesis-free" approaches have increased after the application of computational power to quickly solve multivariate statistical analyses, often in an exploratory fashion. Computerized tracking also allows for the decomposition of animal movement in terms of linear and angular velocity, turning frequency, formation of "homebases," etc. As a result, the number of endpoints derived from a single animal increased dramatically, and some authors exploit that to produce multivariate analyses. For example, Cachat et al. (2011) exploited the three-dimensional character of fish movement and analyzed zebrafish swimming in a novel environment. The combination of manual and automated decomposition of swimming patterns produced 15 endpoints; these were evaluated after "anxiogenic" (repeated morphine withdrawal, alarm pheromone exposure, caffeine, belonging to a highly anxious phenotype) or "anxiolytic" (chronic ethanol, chronic fluoxetine, chronic morphine, acute nicotine) treatments, and the results assessed with hierarchical clustering analysis. The authors revealed that all anxiogenic treatments cluster together, while all anxiolytic treatments cluster together, based on their behavioral effects. Moreover, they also revealed two main behavioral clusters. This exploratory analysis, however, was not followed by confirmatory techniques and therefore should not be interpreted as revealing the underlying structure of zebrafish responses to novelty; moreover, the hypothesis-free character of the approach makes it difficult to understand specifically what each cluster or sub-cluster means in behavioral terms.

Conclusion

Animal models are used to gain insight into studying brain-behavior relations under controlled conditions and to enable predictions about these relations in humans and/or a species other than the one studied or in the same species under conditions different from those under which the study was performed (van der Staay 2006). Animal models need to be thoroughly developed and validated, keeping in mind the intended use of the model. The purposes of the model determine the set of validation criteria that it must fulfill. Although a "classical" set of criteria has been proposed by Willner (Willner 1984, 1986, 1991), new approaches have suggested extended sets of validation criteria (Belzung and Lemoine 2011; Maximino et al. 2010; McOmish et al. 2014; van der Staay et al. 2009). These criteria may also be applied for selecting an (established) animal model for neurobehavioral research.

Cross-References

- ▶ Animal Welfare
- ▶ Behavior Modification
- ▶ Experimental Psychology
- ▶ Transgenic

References

Atanasova, N. A. (2015). Validating animal models. *Theoria, 30*(2), 163–181. https://doi.org/10.1387/theoria.12761.

Belzung, C. (2014). Innovative drugs to treat depression: Did animal models fail to be predictive or did clinical trials fail to detect effects? *Neuropsychopharmacology Reviews, 39*(5), 1–10. https://doi.org/10.1038/npp.2013.342.

Belzung, C., & Lemoine, M. (2011). Criteria of validity for animal models of psychiatric disorders: Focus on anxiety disorders and depression. *Biology of Mood & Anxiety Disorders, 1*(9), 14. https://doi.org/10.1186/2045-5380-1-9.

Burrows, E. L., McOmish, C. E., & Hannan, A. J. (2011). Gene–environment interactions and construct validity in preclinical models of psychiatric disorders. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 35*(6), 1376–1382. https://doi.org/10.1016/j.pnpbp.2010.12.011.

Cachat, J., Stewart, A., Utterback, E., Hart, P., Gaikwad, S., Wong, K., . . . Kalueff, A. V. (2011). Three-dimensional neurophenotyping of adult zebrafish behavior. *PLoS One 6*(3), e17597, 14. https://doi.org/10.1371/journal.pone.0017597.

Crabbe, J. C., & Morris, R. G. M. (2004). Festina lente: Late-night thoughts on high-throughput screening of

mouse behavior. *Nature Neuroscience, 7*(11), 1175–1179. https://doi.org/10.1038/nn1343.

Cuthbert, B. N. (2014). The RDoC framework: Facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry, 13*(1), 18–35. https://doi.org/10.1002/wps.20087.

Degeling, C., & Johnson, J. (2013). Evaluating animal models: Some taxonomic worries. *Journal of Medicine and Philosophy, 38*, 91–106. https://doi.org/10.1093/jmp/jht004.

Denayer, T., Stöhr, T., & van Roy, M. (2014). Animal models in translational medicine: Validation and prediction. *New Horizons in Translational Medicine, 2*. https://doi.org/10.1016/j.nhtm.2014.08.001.

Ferreira, G. S., Veening-Griffioen, D. H., Boon, W. P. C., Moors, E. H. M., Gispen-de Wied, C. C., Schellekens, H., & van Meer, P. J. K. (2019). A standardised framework to identify optimal animal models for efficacy assessment in drug development. *PLoS One*, *14*(6), e0218014, 17. https://doi.org/10.1371/journal.pone.0218014.

Gamzu, E. (1985). Animal behavioral models in the discovery of compounds to treat memory dysfunction. *Annals of the New York Academy of Sciences, 444*, 370–393. https://doi.org/10.1111/j.1749-6632.1985.tb37602.x.

Gould, T. D., & Gottesman, I. I. (2006). Psychiatric endophenotypes and the development of valid animal models. *Genes, Brain and Behavior, 5*(2), 113–119. https://doi.org/10.1111/j.1601-183X.2005.00186.x.

Gouveia, A., Jr., & de Brito, T. M. (2015). Animal models of psychopathology and its relation to clinical practice. In P. A. Gargiulo & H. L. M. Arroyo (Eds.), *Psychiatry and neuroscience update*. https://doi.org/10.1007/978-3-319-17103-6_22.

Greek, R., & Rice, M. J. (2012). Animal models and conserved processes. *Theoretical Biology and Medical Modelling 9,* 40, 33. https://doi.org/10.1186/1742-4682-9-40.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science, 70*, 1195–1205. https://doi.org/10.1086/377400.

Hoffman, K. L. (2016). New dimensions in the use of rodent behavioral tests for novel drug discovery and development. *Expert Opinion on Drug Discovery, 11*(4), 343–353. https://doi.org/10.1517/17460441.2016.1153624.

Kas, M. J. H., Fernandes, C., Schalkwyk, L. C., & Collier, D. A. (2007). Genetics of behavioural domains across the neuropsychiatric spectrum; of mice and men. *Molecular Psychiatry, 12*, 324–330. https://doi.org/10.1038/sj.mp.4001979.

Kas, M. J. H., Gelegen, C., Schalkwyk, L. C., & Collier, D. A. (2009). Interspecies comparisons of functional genetic variations and their implications in neuropsychiatry. *American Journal of Medical Genetics – Part B, 150B*, 309–317. https://doi.org/10.1002/ajmg.b.30815.

Kas, M. J. H., Krishnan, V., Gould, T. D., Collier, D. A., Olivier, B., Lesch, K.-P., et al. (2011). Advances in multidisciplinary and cross-species approaches to examine the neurobiology of psychiatric disorders. *European Neuropsychopharmacology, 21*, 532–544. https://doi.org/10.1016/j.euroneuro.2010.12.001.

LaPorte, J. L., Ren-Patterson, R. F., Murphy, D. L., & Kalueff, A. V. (2008). Refining psychiatric genetics: From 'mouse psychiatry' to understanding complex human disorders. *Behavioural Pharmacology, 19*, 377–384. https://doi.org/10.1097/FBP.0b013e32830dc09b.

Maximino, C., & van der Staay, F. J. (2019). Behavioral models in psychopathology: Epistemic and semantic considerations. *Behavioral and Brain Functions 15*, 1, 11. https://doi.org/10.1186/s12993-019-0152-4.

Maximino, C., Marques de Brito, T., & Gouveia, A., Jr. (2010). Construct validity of behavioral models of anxiety: Where experimental psychopathology meets ecology and evolution. *Psychology & Neuroscience, 3*(1), 117–123. https://doi.org/10.3922/j.psns.2010.1.015.

Maximino, C., da Silva, A. W. B., Araújo, J., Gomes Lima, M., Miranda, V., Puty, B., . . . Herculano, A. M. (2014). Fingerprinting of psychoactive drugs in zebrafish anxiety-like behaviors. *PLoS One 9*(7), e103943, 8. https://doi.org/10.1371/journal.pone.0103943.

McKinney, W. T., Jr., & Bunney, W. E., Jr. (1969). Animal model of depression, I. Review of evidence: Implications for research. *Archives of General Psychiatry, 21*, 240–248. https://doi.org/10.1016/B978-0-08-023725-1.50013-6.

McNaughton, N., & Zangrossi, H., Jr. (2008). Theoretical approaches to the modeling of anxiety in animals. In R. J. Blanchard, D. C. Blanchard, G. Griebel, & D. Nutt (Eds.), *Handbook of anxiety and fear* (Vol. 17, pp. 11–27). https://doi.org/10.1016/S1569-7339(07)00002-1.

McOmish, C. E., Burrows, E. L., & Hannan, A. J. (2014). Identifying novel interventional strategies for psychiatric disorders: Integrating genomics, 'enviromics' and gene–environment interactions in valid preclinical models. *British Journal of Pharmacology, 171*, 4719–4728. https://doi.org/10.1111/bph.12783.

Stewart, A. M., & Kalueff, A. V. (2015). Developing better and more valid animal models of brain disorders. *Behavioural Brain Research, 276*, 28–31. https://doi.org/10.1016/j.bbr.2013.12.024.

Stewart, A. M., Nguyen, M., Poudel, M. K., Warnick, J. E., Echevarria, D. J., Beaton, E. A., et al. (2015). The failure of anxiolytic therapies in early clinical trials: What needs to be done. *Expert Opinion on Investigational Drugs, 24*(4), 543–556. https://doi.org/10.1517/13543784.2015.1019063.

van der Staay, F. J. (2006). Animal models of behavioral dysfunctions: Basic concepts and classifications, and an evaluation strategy. *Brain Research Reviews, 52*(1), 131–159. https://doi.org/10.1016/j.brainresrev.2006.01.006.

van der Staay, F. J., Arndt, S. S., & Nordquist, R. E. (2009). Evaluation of animal models of neurobehavioral

disorders. *Behavioral and Brain Functions 5*, 11, 23. https://doi.org/10.1186/1744-9081-5-11.

van der Staay, F. J., Arndt, S. S., & Nordquist, R. E. (2014). Developing mouse models of neurobehavioral disorders: When is a model a good model? In S. Pietropaolo, F. Sluyter, & W. E. Crusio (Eds.), *Behavioral genetics of the mouse* (pp. 3–17). https://doi.org/10.1007/CBO9781107360556.

Willner, P. (1984). The validity of animal models of depression. *Psychopharmacology, 83*, 1–16. https://doi.org/10.1007/BF00427414.

Willner, P. (1986). Validation criteria for animal models of human mental disorders: Learned helplessness as a paradigm case. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 10*, 677–690. https://doi.org/10.1016/0278-5846(86)90051-5.

Willner, P. (1991). In P. Willner (Ed.), *Behavioural models in psychopharmacology* (pp. 3–18). Cambridge: Cambridge University Press.

Wright, C. D. (2002). Animal models of depression in neuropsychopharmacology qua Feyerabend philosophy of science. In S. P. Shodov (Ed.), *Advances in psychologyy research* (Vol. 13, pp. 129–148). New York: NovaScience Publishers.