Section 1 **Introduction to genetic mouse models of neurobehavioral disorders**

Chapter 1

Introduction

Genetic mouse models of neuropsychiatric disorders

Frans Sluyter, Susanna Pietropaolo, and Wim E. Crusio

Animal modeling is an interactive process in which one or more species (the "models") are studied to gain insight on a trait or disorder in humans including the testing of (new) hypotheses and (pharmaco)therapies. Roughly speaking, a model's usefulness depends on the strength of its validity, which has three components: face, predictive, and construct validity. These, in turn, depend on the information we have about the trait or disorder. Thus, the more we know about the cause(s), genetic and/or environmental, and the exact pathophysiology of a disorder, the better we are able to model certain aspects of that disorder. In recent decades, mice have replaced rats to become the animal model of choice for behavioral neuroscientists. The most important reason for this has been the rapid advances in genetic engineering over the last two decades, to the extent that there is now a wealth of distinct mouse techniques available to mimic (or test hypotheses on) the pathophysiology of a disorder. In addition, both the mouse brain and genome are similar to those of humans.

Psychiatric disorders are among the most fascinating human diseases as they touch directly on that which makes us human: our minds. Whereas, say, a heart patient may be sick and suffering, cardiac disease touches the mind only indirectly (by the stress it generates, for example) and the afflicted patient remains recognizably the same person. Not so with many psychiatric disorders, which not only can be life-long debilitating diseases, but directly affect and in some cases dramatically change a patient's mind. Modeling a disordered mind and its consequent behavior, however, can be very challenging, because it is difficult to develop animal tests that convincingly and consistently mimic human symptoms. In addition, for most psychiatric disorders we lack objective and reliable information about their etiology. Psychiatric diagnoses are, to a great extent, subjective and based on the presence of a minimal number of symptoms from a list of symptoms during a certain period of time (DSM-5, 2013). For instance, a diagnosis of depression, or major depressive disorder in DSM terms is based on the presence of five symptoms out of a list of nine (DSM-5, 2013), which means that theoretically two persons with the same diagnosis may share only one symptom. This heterogeneity is corroborated even further by the fact that the diagnostic criteria for depression are partly shared with anxiety disorders and that one single episode of mania changes the diagnosis to bipolar disorder, which is presumably a distinct pathophysiological entity (Krishnan and Nestler, 2008). It is therefore not surprising that the search for genes (or DNA markers) underlying (or reliably associated with) depression has been largely disappointing as opposed to, for example, the recently published list of genetic markers for hormonally mediated cancers, for which objective and reliable biomarkers exist. (See Sakoda et al., 2013 for a commentary on the dozen high-impact papers reporting over 70 new susceptibility loci for breast, ovarian, and prostate cancers.) In addition to the lack of objective biological markers and variation in symptoms, the impossibility of modeling typically human symptoms such as guilt and suicidal ideation raises another barrier in modeling depression. Consequently, most models of depression, including genetic mouse models, basically test hypotheses about the disorder (e.g., by changing the underlying genetics of a neurobiological pathway known to be involved in a subset of affected individuals, see also Chapter 22 for a critical assessment of mouse models for depression).

There are exceptions, though. The genetic causes of neurodevelopmental disorders such as Fragile X or Rett syndrome are known and these disorders can be reliably modeled using genetically engineered mice, i.e., *Fmr1* and *Mecp2* knockout (KO) mice, respectively. These models have high construct validity as they capture the essence of Fragile X and Rett syndrome and can be studied invasively – an (ethical) impossibility in humans – to learn about the pathophysiology underlying these disorders and to search for suitable treatments. For instance, brain Rho GTPases have been identified as an innovative therapeutic target in *Mecp2* knockouts and the administration of cytotoxic necrotizing factor 1 (CNF1, which activates Rho GTPases) has been shown to markedly improve Rett symptomatology in these mice (see Chapter 13). Similarly, *Fmr1*-KO mice have been employed to design pharmacological and non-pharmacological therapeutic approaches, some already leading to clinical trials (see Chapter 14).

Genetic mouse models are also helpful in modeling the effects of rare genetic variants. An excellent example hereof is

Behavioral Genetics of the Mouse: Volume II. *Genetic Mouse Models of Neurobehavioral Disorders*, eds. S. Pietropaolo, F. Sluyter, and W. E. Crusio. Published by Cambridge University Press.

the study of Bevilacqua et al. (2010), who discovered an association between impulsivity and the gene coding for the 5-HT2B receptor (*HTR2B)* in a Finnish subpopulation. Violent offenders whose crimes were characterized by a high degree of impulsivity were statistically more likely to lack functional 5-HT2B receptors as a result of a mutated *HTR2B* gene. Bevilacqua et al. (2010) subsequently generated *Htr2b* knockout mice and tested these animals for five separate measures of impulsivity and novelty seeking. They found that knockouts were more active in a novel environment, displayed an increased number of contacts with a novel object, and were less likely to wait for a larger but later reward, a behavioral profile similar to the genetically affected violent offenders.

In fact, some think that the effects of individual mutations on the pathogenesis of psychiatric disorders may be more important than hitherto thought. Thus, based on recent findings from whole-exome and whole-genome sequencing, Mitchell et al. (2011) postulate that psychiatric disorders are actually umbrella terms for large numbers of distinct genetic disorders that happen to result in similar spectra of symptoms. They further propose to capture these (*de novo*) mutations (which may also include duplications and translocations) in mouse models with direct construct validity, i.e., where the genetic manipulation results in a defect homologous to the actual cause of the condition in humans. These "direct" animal models of genetic etiology can then be further analyzed using the full arsenal of modern behavioral neuroscience. Insel (2007) calls these types of animal models "model animals," making a careful distinction between models that phenotypically resemble aspects of mental disorders (old-fashioned animal models) and models with the molecular and cellular abnormalities found in mental disorders (model organisms).

However, the prevailing opinion regarding the pathogenesis of neurobehavioral disorders is still the polygenic/threshold model in which what is inherited is not so much a disorder as a liability to a disorder contributed to by multiple genetic and environmental effects. Each of these, by themselves, would only have a small effect on risk, but when the collective burden of such alleles passes a putative threshold, the system would be pushed into a pathogenic state. The polygenic/threshold model is more about probability as opposed to the mutation model, which is more about causality. Moreover, and perhaps more importantly from an animal modeling point of view, the relatively small contribution of each effect makes it difficult to find genetic disease variants and construct appropriate models. The result is that for most neurobehavioral disorders, precise genetic information is either lacking or not very reliable. Consequently, in this framework genetic models are either speculative or only capture a small part of the underlying etiology. As for the speculative side of genetic modeling, Nestler and Hyman (2010) call this "reversing the direction of validation," in which observed pathology in genetic (mouse) models may be sought in human patients, either in postmortem tissue or non-invasive imaging.

Ideally, genetic mouse modeling is a two-way street where human (liabilities to) pathologies, either on a genetic or on a neuro-circuitry level, are mirrored in model animals, which, in turn, inform and steer human studies. As long as we are clear and honest about what we (attempt to) model and as long as we keep the limitations of modeling in mind, genetically engineered mice can be very effective in elucidating the pathophysiology of neurobehavioral disorders and ultimately in finding successful (pharmaco)therapies. Last but not least, although the vast majority of genetic mouse models presented in this volume are the result of active gene (or chromosome) engineering, we should not forget about the traditional genetic mouse models, i.e., artificial selection lines and inbred strains, which still have added value in understanding and modeling neurobehavioral disorders. An outstanding illustration hereof is the work of Phillips et al. (Chapter 27) who used a variety of short-term bidirectionally selected lines to gain insight into the neurobiology of amphetamine addiction.

Although this book presents quite a few success stories where genetic mouse models have been very effective in elucidating disease mechanisms, it would not be fair to skip the equally numerous failures. For example, sometimes a mutation with a dramatic effect in humans has much more moderate effects in mice. An example of this is the *Fmr1* KO mouse (Chapter 14). Although this animal has the same molecular defect as human patients with Fragile X syndrome (i.e., no fragile X mental retardation protein (FMRP) expression) and does display many of the same symptoms that human patients show, the severity of the disorder is much reduced in mice.

Another problem is the rather frequent failure to replicate findings obtained in different laboratories. Ever since the landmark study by Crabbe et al. (1999), this is often brushed away as being unavoidable variation due to interlaboratory differences. This is not the complete truth for several reasons. First of all, it is often overlooked that the Crabbe et al. study actually showed that many behavioral differences *can* be reliably reproduced in different laboratories and this often over decades (Wahlsten et al., 2006). Second, and in our opinion even more seriously, we feel that many failures to replicate are due to conceptual inadequacies in our arsenal of behavioral tests. Many tests have never been properly studied and validated. For many other tests, validation has been only cursory, testing just two groups of animals, one of them treated with some pharmacologically active substance of supposedly known effect and the other the controls. It is becoming increasingly clear that tests that purportedly measure the same behavioral quality often give divergent results even in the same lab and in the hands of the same experimenters. An example is the study of Mineur et al. (2006), who tested animals from different inbred strains in both the Porsolt forced swim test and the tail suspension test. Although both tests are supposed to measure the very same behavioral construct, namely depression-like behavior ("behavioral despair"), the results of both tests were dramatically

different. Thus, it would appear that the refinement and finesse of our current behavioral tools do not match those of our genetic tools. Improving our understanding of our behavioral methods will therefore be an important challenge in the near future for neuroscience, and neurogenetics in particular.

We would like to finish this introduction on a more optimistic note, however. The past two decades have shown the power of the "new genetics" (now also including more neurocircuitry-focused techniques such as optogenetics) to generate mice that are genetically tailored to suit the needs of behavioral neurogeneticists wishing to model a human neuropsychiatric disorder. And despite some unavoidable problems, this volume presents numerous examples of the power of this approach, leading to important insights into the mechanisms underlying these fascinating ailments, with potentially significant benefits for human well-being.

References

Bevilacqua, L., Doly, S., Kaprio, J., et al. (2010) A populationspecific HTR2B stop codon predisposes to severe impulsivity. *Nature* 468: 1061–1066.

Crabbe, J.C., Wahlsten, D., and Dudek, B.C. (1999) Genetics of mouse behavior: interactions with laboratory environment. *Science* 284: 1670–1672.

DSM-5 (2013) *Diagnostic and Statistical Manual of Mental Disorders, 5th edn (DSM-5)*. Washington, DC: American Psychiatric Association.

Insel, T.R. (2007) From animal models to model animals. *Biol Psychiatry* 62: 1337–1339.

Krishnan, V. and Nestler, E.J. (2008) The molecular neurobiology of depression. *Nature* 455: 894–902.

Mineur, Y.S., Belzung, C., and Crusio, W.E. (2006) Effects of unpredictable chronic mild stress on anxiety and depression-like behavior in mice. *Behav Brain Res* 175: 43–50.

Mitchell, K.J., Huang, Z.J., Moghaddam, B., and Sawa, A. (2011) Following the genes: a framework for animal modeling of psychiatric disorders. *BMC Biology* 9: 76.

Nestler, E.J. and Hyman, S.E. (2010) Animal models of neuropsychiatric disorders. *Nat Neurosci* 13: 1161–1169.

Sakoda, L.C., Jorgenson, E., and Witte, J.S. (2013) Turning of COGS moves forward findings for hormonally mediated cancers. *Nat Genet* 45: 345–348.

Wahlsten, D., Bachmanov, A., Finn, D.A., and Crabbe, J.C. (2006). Stability of inbred mouse strain differences in behavior and brain size between laboratories and across decades. *Proc Natl Acad Sci* 103: 16364–16369.

Section 1 Introduction to genetic mouse models of neurobehavioral disorders

Chapter 2

Developing mouse models of neurobehavioral disorders

When is a model a good model?

F. Josef van der Staay, Saskia S. Arndt, and Rebecca E. Nordquist

Animal models in neurobehavioral research

To make sense of a discussion of animal models, one first has to understand both the purpose of such models, and their benefits to research, as well as the limitations on their interpretation.

(McMillen, 1997: 409)

In line with the above statement, a model is a good model when it serves its purpose (Geyer and Markou, 1995) and advances scientific insight. This prompts a number of questions: What is a model? What is the purpose of a model? How is a model developed and validated? How can we evaluate a model, i.e., decide whether this is indeed a good model? We will try to address these questions, with emphasis on model evaluation.

We define animal models in the behavioral neurosciences, which include models of neurobehavioral disorders, as follows:

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: 133–134)

Purpose of animal models

Animal models are developed for a specific purpose (Festing, 2004; Holmes, 2003; Massoud et al., 1998). For example, animal models of neurobehavioral disorders are used to enhance our understanding of their underlying substrates and mechanisms. The relation between brain and behavior can be investigated experimentally by using pharmacological agents, lesions, or animals with naturally occurring or experimentally induced deficits to distinguish between processes, subprocesses, and modulating influences (Cernak, 2005; D'Mello and Steckler, 1996). Of particular interest is the identification of new targets, pathways, and mechanisms of drug action (Matthews and Kopczynski, 2001; Snaith and Törnell, 2002; West et al., 2000).

Animal models simplify complex phenomena, but at the same time the use of an animal model should allow the confirmation and/or rebuttal of specific hypotheses (Marcotte et al., 2001). If the animal model is too complex to provide clearer answers than other methods, then its availability and application does not advance scientific insight and it is not useful (Massoud et al., 1998). However, if ethical considerations prevent experimental manipulation of the target species, e.g., humans, then it may be "permissible" to use phylogenetically "lower" species in animal models to gain information.

Animal models can also be used to translate insights gained in preclinical animal studies to the clinical setting (and vice versa; Porges, 2006; Waldman and Terzic, 2010). For instance, animal models can be used to assess the effects of putative neuroprotective, antidegenerative, revalidation-supporting, mental health-promoting, and/or cognition-enhancing compounds or treatments (Allain et al., 1998; Frazer and Morilak, 2005; Hitzemann, 2000; Willner, 1998; Wong et al., 2002), and to evaluate the risks (safety, teratology, toxicology) associated with these treatments (Bolon, 2004; Cavero, 2010).

Validity of animal models

Nearly three decades ago, Willner (1986) argued that animal models should possess three types of validity: *face validity, predictive validity*, and *construct validity* (Figure 2.1), a categorization that has since been adopted by many researchers (e.g., Chesselet and Richter, 2011; Homberg, 2013). External validity, i.e., the degree of generalizability of experimental results obtained in the laboratory to the "outside world," has since been added to this list (Guala, 2003). Others have reduced or extended the types of validity that a model should possess (Belzung and Lemoine, 2011; Cryan and Sweeney, 2011; Tordjman et al., 2007; Young et al., 2010). It should be noted that the validity of a model is not a measure of the truth of findings obtained with the model (Massoud et al., 1998).

Behavioral Genetics of the Mouse: Volume II. *Genetic Mouse Models of Neurobehavioral Disorders*, eds. S. Pietropaolo, F. Sluyter, and W. E. Crusio. Published by Cambridge University Press.

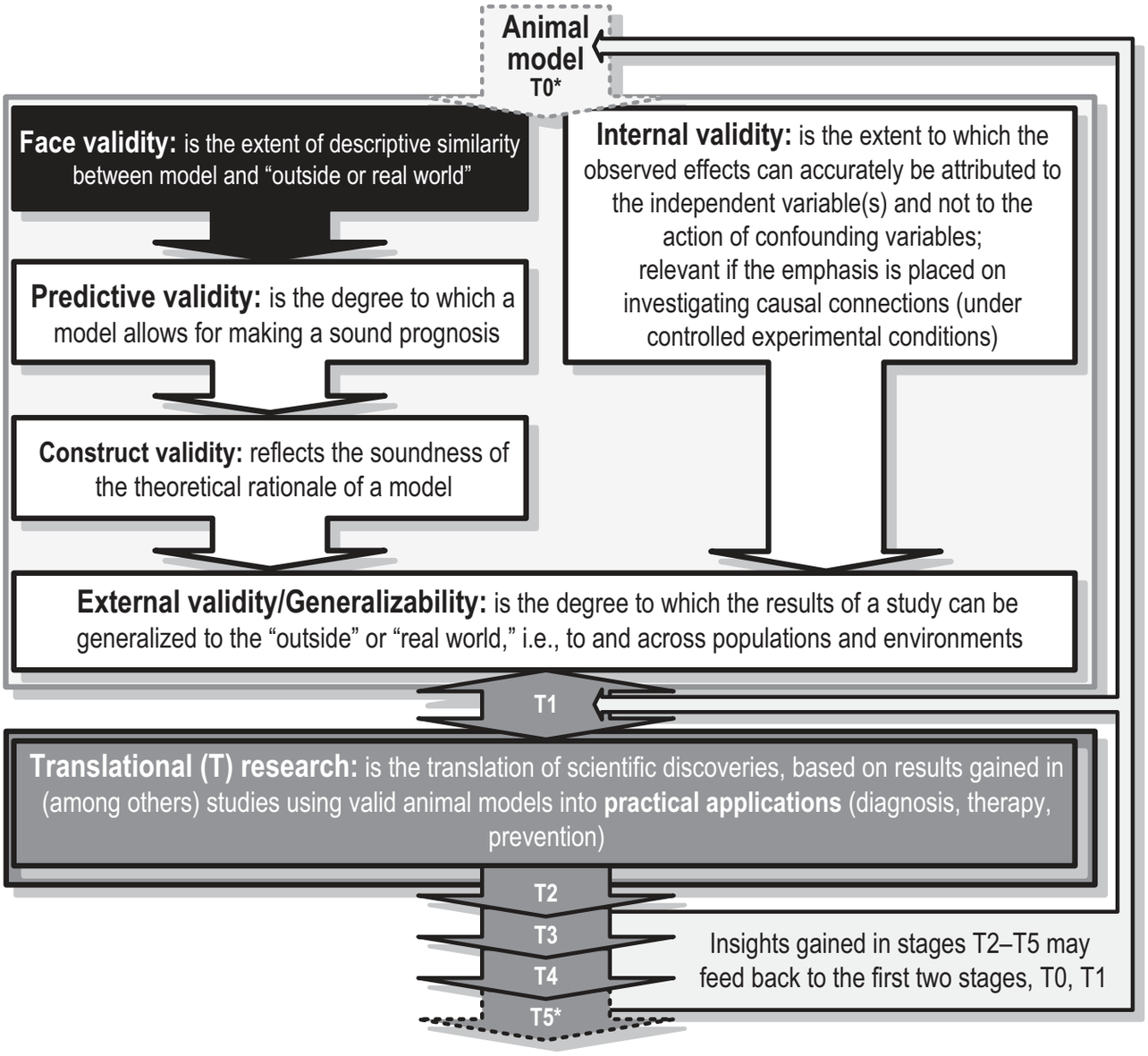

Figure 2.1 Hierarchy of validities in animal models, and stages in translational research.
The upper left column shows the hierarchy of validities that is taken as the basis for animal model development. Face validity is in a special position, as a lack of face validity does not *per se* invalidate an animal model. Validity is subdivided into two classes: internal and external validity. This differentiation is only applicable to experiments that investigate causal relationships (modified and extended from Fig. 10.1 in van Zutphen et al., 2009: 201).
Translational research distinguishes between different stages, most commonly T1–T4 (Drolet and Lorenzi, 2011; Waldman and Terzic, 2010), of which T1 is "the transfer of new understandings of disease mechanisms gained in the laboratory into the development of new methods for diagnosis, therapy, and prevention and their first testing in humans" (Woolf, 2007).
*: Waldman and Terzic (2010) suggested extensions of translational stages to include T0, preclinical research (*in vitro* and *in vivo* animal model-based research), and T5, improving the wellness of populations by reforming suboptimal social structures.
The bidirectional and recursive relationship between animal models, translation to applications, and reverse translation to animal models is indicated in Figure 2.1 by the two-headed arrow to the first stage of translational research (T1). However, insights gained in later translational stages may also feed back to earlier stages, including T0 and T1.
(With kind permission from Reed Business Education)

Models are validated to increase confidence in the model. Validation provides information about the plausibility and consistency of the interpretation of data generated with the animal model. Validity is a major criterion for establishing the worth of animal models (Holmes, 2003), although it should be recognized that no animal model is valid in all situations and for all purposes. Validity is restricted to a specific use of the model, and thus must always be open for discussion and re-evaluation (Silva, 1993).

Face validity

Face validity is the degree of descriptive similarity between, for example, the behavioral dysfunction seen in the animal model and the symptoms caused by a particular neurobehavioral disorder in humans. Some researchers hold that a model has face validity if it mimics the fundamental (behavioral) characteristics of the disease or symptomatology being modeled (Wickens et al., 2011). However, there is no consensus on the importance of face validity, with some investigators considering face validity to be a major (or even the most important, e.g., Holmes, 2003) criterion for model evaluation, whereas others consider it incorrect to put too much emphasis on this criterion (e.g., Sarter and Bruno, 2002; Wang et al., 2009). In fact, face validity may prove to be an unrealistic criterion (Matthews et al., 2005; Wang et al., 2009) and may form a barrier to the development of animal models using phylogenetically lower animal species (Burne et al., 2010; Sufka et al., 2006). In general, the likelihood of symptoms being similar (Willner, 1986) is higher in species that are phylogenetically closer to humans (see comments by Bezard, 2006).

Predictive validity

An animal model possesses predictive validity if it allows extrapolation of the effect of a particular experimental manipulation from that species to another species, including humans, and from one condition (e.g., the laboratory) to another (e.g., the "real world"), or from one time point to another (see also the definition of "animal model" given above). Predictive validity may have some components in common with external validity (or generalizability; see below).

The concept of predictive validity is often used in a narrower sense in psychopharmacological studies (e.g., Bourin et al., 2001; Cryan and Slattery, 2007; Sarter et al., 1992; Whiteside et al., 2008), to indicate the ability of drug screening or an animal model to correctly identify the efficacy of a putative therapeutic agent (Wright, 2002), usually with the most effective treatment currently available serving as "gold standard." However, for some diseases only a few (weakly) effective therapeutic agents are available, and sometimes there are no clinically active drugs (Markou et al., 2009), which makes it impossible to establish the predictive validity of animal models.

Construct validity

Construct validity reflects the soundness of the theoretical rationale (Wright, 2002), i.e., the degree of fit between the theoretical rationale and the true nature of the symptoms/syndrome to be mimicked (Holmes, 2003). Construct validity expresses how well the manipulations (independent variables) and the measurements (dependent variables) correspond with the theoretical hypotheses to be tested (Lubow, 2005). It is a theory-driven, experimental substantiation of the behavioral, pathophysiological, and/or neuronal components of the model (Sarter and Bruno, 2002). Sarter and Bruno (2002) argue that construct validity is the most important criterion for animal models because it addresses the soundness of the theory underlying the model, and because it provides the framework for interpreting data generated by the model. However, this implies that there are comprehensive theories about the brain–behavior relation of the phenomenon under investigation, which is not the case if little is known about the underlying pathophysiological conditions of the disease or dysfunction (Einat, 2011).

Generalizability

Assessment of the generalizability (or "external validity") of experimental findings should be an integral part of experimental work in general and should thus be an integral part of model development. This empirical process can be performed by *systematic replications* or *differentiated replications*, i.e., replications of the original studies in which a particular set of independent variables is varied systematically. This process can be used to evaluate whether the results obtained are robust across, for example, rearing and housing conditions, ages, gender, and test conditions or tests used (Figure 2.2). Ideally, a replication study is not merely a repetition of an earlier study, but extends the scope of previously performed studies, allowing statements to be made about the generality of results (Lindsay and Ehrenberg, 1993; van der Staay, 2010).

Internal versus external validity

Alternatively, one may evaluate the internal and external validity of an animal model (see Figure 2.1). It is generally accepted that the measures taken to increase internal validity may compromise external validity by restricting the range of conditions under which the relationship between dependent and independent variables is tested. That said, a higher internal validity fosters higher explanatory power (Eifert et al., 1999).

Internal validity here refers to the quality of the experimental evaluation of the animal model, i.e., to how well a study was performed, how strictly putative confounding variables were controlled (see, e.g., Schellink et al., 2010; see also Figure 2.2), and how confident one can be that the changes observed in the dependent variable(s) are caused by experimental manipulation of the independent variable(s), and not by confounders, i.e., factors that might also affect the independent variable and may offer alternative explanations for the results obtained. It is essential to establish that results are valid within the laboratory setting (internal validity; Guala, 2003; Kazdin and Rogers, 1978) before speculating about whether findings can be generalized to other settings. Consequently, internal validity requires that the experiment is highly standardized, that the subjects included in the study represent a homogeneous sample, and that putative interfering variables and extraneous influences are strictly controlled in order to increase the sensitivity to detect causal relationships (if they exist) (Kazdin and Rogers, 1978). Internal validity addresses the question whether the

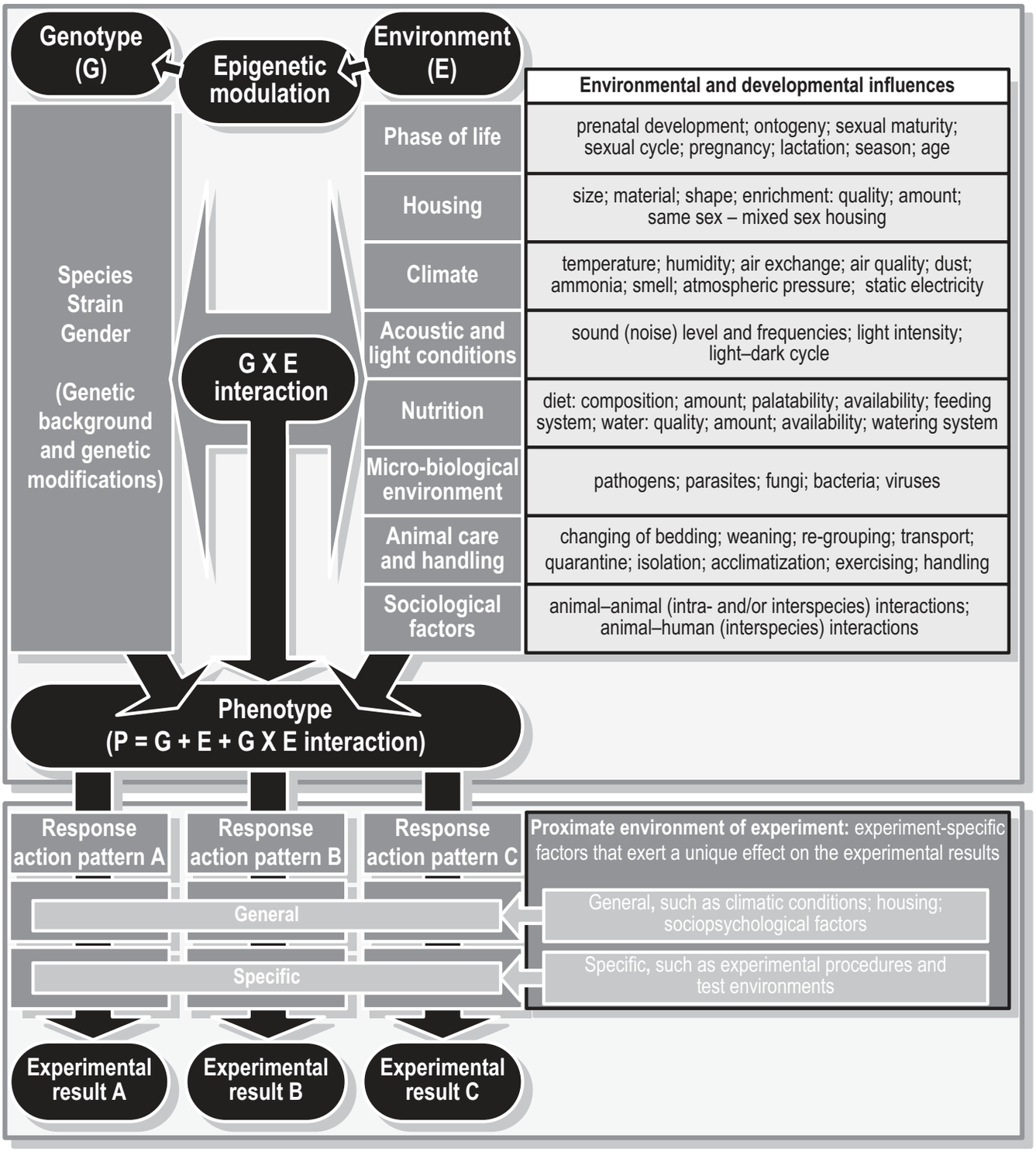

Figure 2.2 The genotype, environmental (and developmental) influences and their interactions determine the phenotype (upper part of the figure inspired by Öbrink and Rehbinder, 2000). In the lower part, the concept of "response action pattern" is schematically presented (modified from Fig. 1 in van der Staay et al., 2010 with permission from Wiley). General and specific proximal factors can affect the results of an experiment, even if strictly defined and controlled phenotypes are used, i.e., the housing and testing conditions might modulate the experimental results.

experimental intervention makes a difference, whereas external validity addresses the question whether an experimental effect can be generalized.

External validity (or *generalizability*) is the extent to which the results obtained using a particular animal model can be generalized/applied to and across populations (and ultimately species) (Mace, 1996) and environments, or "the extent to which experimental findings make us better able to predict real-world behavior" (Mook, 1989). Assessment of the external validity is an empirical process. The external validity of animal models may benefit from multiple-site testing, using output parameters of proven clinical relevance, including animals raised and kept in different environments, or using animals in different health states (Jucker, 2010).

Distinction between generalizable and translational animal models

The use of validated animal models to develop novel therapeutics may be considered translational research, where translation refers to the process in which knowledge generated in one area of research is applied in another area of research in order to advance goals in that area (Abernethy and Wheeler, 2011). Animal models are the first step in the translational continuum that has been described as consisting of three (Abernethy and Wheeler, 2011; Drolet and Lorenzi, 2011), four (Lander and Atkinson-Grosjean, 2011), or five (Waldman and Terzic, 2010) distinct stages (see also Trochim et al., 2011). For our purposes, only the first translational stage (T1) is of relevance (see Figure 2.1). Concerns have been raised about the quality and relevance of animal models to translational research and about the appropriate choice of an animal model (e.g., Plath et al., 2011; Pratt et al., 2012). These concerns are an obstacle to the successful translation of findings obtained with animal models to their practical application (Sabroe et al., 2007).

Translational animal models are a subset of generalizable animal models. "Translational value" and "translational relevance" are obviously concepts that bear similarity with the concept of "predictive validity" or may even be considered as basically the same concept, although with a clear focus on applicability (diagnosis, therapy, and prevention). The predictive validity of animal models can be tested experimentally, by comparing the effects of a therapeutic in a model with its effect in patients. As with predictive validity, the translational value or translational relevance of a model can only be determined retrospectively, i.e., by proving that the insights derived from an animal model could successfully be translated to practical applications. The probability of successful translation of preclinical, animal research-based findings to the clinic may be increased by using models with proven construct validity (Kimmelman and London, 2011; Pratt et al., 2012), by using a broader range of relevant models (Pratt et al., 2012), e.g., including animals showing comorbidity, applying a treatment regimen that more closely matches clinical practice (Green et al., 2011), and by applying appropriate behavioral tests that can distinguish between different behavioral domains and (endo)phenotypes (Homberg, 2013).

The term "reverse translational" has recently been coined, i.e., the translation of clinical observations to basic research (Weston et al., 2010), a process of induction. Holschneider and colleagues (2011), for example, performed "reverse translational" research in which they used a rodent model that expresses the homologue of an endophenotype identified in patients, and characterized it pharmacologically. Sinha et al. (2011) used the "reverse translational" approach to assess whether neuropharmacological findings in humans can be used to investigate underlying mechanisms in an appropriate animal model. Reverse translational approaches thus may contribute to the development of new animal models. Translationally relevant animal models need to be developed (Pratt et al., 2012), based on an intensive interaction of animal research scientists and clinical researchers (Markou et al., 2009).

The criteria that translational animal models need to fulfill may be different from those of animal models intended for use in basic research, i.e., additional criteria may be relevant if the model is a component of the translational T1 stage. The key definitions per stage may differ for different classifications of the translational continuum, but they all describe translational research as a process that moves from basic/preclinical research (which includes the use of animal models) to the clinical application of findings, and finally, the impact on public health. This process is bidirectional (Trochim et al., 2011).

Assessing reliability/replicability and external validity/generalizability

The replicability of results is a fundamental aspect of science (Kelly, 2006; Muma, 1993; Park, 2004). Experimental results are considered preliminary as long as they have not been corroborated, and preferably by investigators other than those who originally performed the investigations (Levin, 1998; Rosenthal, 1991; van der Staay, 2006). Replication is essential for determining the reliability and generalizability of study findings. This is particularly important with newly developed genetic models, where the original study may have limited statistical power because of the small number of animals used, and where successful replication increases confidence in the results and implications of findings obtained with the model (Palmer, 2000).

One may apply a "replication battery" to estimate the reliability/replicability and generalizability of the results of the first, original study. This replication battery can be conceived as a two-, or eventually multiple-tiered, experimental approach and may comprise close, partial, or extended replications (Kelly, 2006; discussed extensively by van der Staay et al., 2010). Extended replications make it possible to identify the conditions under which generalization does not hold, and they help to detect putative confounding variables and to assess their effects (Lindsay and Ehrenberg, 1993). These replications expose the strengths and weaknesses of findings and the limits of their

generalizability. Extended replications can generate new insights that may initiate a new iterative cycle of generating revised or new hypotheses and subsequently hypothesis-testing studies (Townsley and Johnson, 2008).

Animal welfare, minimized discomfort: ethical considerations

The number of potential mouse models of neurobehavioral disorders is constantly increasing. Genetically modified mice may offer a highly relevant disease model as a consequence of targeting the disease or a subset of specific symptoms and/or disease pathogenesis (Brown and Murray, 2006). The aim of these models is to generate signs similar to those seen in the target species (usually humans) (Doyle et al., 2012). As a consequence, the animals used to model neurobehavioral disorders may experience discomfort (le Bars et al., 2001; Mertens and Rülicke, 2007; van Zutphen and De Deynyn, 2000). For this reason, animal welfare should be closely monitored during the different stages of model development, in colonies of specific strains, and in genetically engineered animals.

Welfare issues that arise during the development of a genetically altered line include health problems, increased lethality, and a decrease in adaptive fitness (e.g., Gerrits et al., 2008; Joubert et al., 2012), which in turn might influence fetal mortality and fetal welfare in breeding programs (Mellor, 2010). In some strains, poor mothering may necessitate cross-fostering (i.e., the foster mother's newborns must be removed). Moreover, animals that do not express the desired genotype are frequently considered as surplus animals that will not be used in experimental studies (except if they can serve as controls). Separate lines of mutant mice may be necessary to allow the breeding of model animals carrying multiple mutations, which can substantially increase the number of surplus animals. This contravenes the principle of "reduction" in the 3Rs ("replacement, reduction, and refinement"; Manciocco et al., 2009), whereas the development of models based on genetically modified animals is generally considered a refinement, i.e., the 3Rs may be conflicting (Fenwick et al., 2009).

Careful evaluation is needed to determine whether the observed dysfunctions and associated discomfort are part of the phenotype under consideration, or whether steps should be taken to reduce discomfort. Compromised animal welfare may interfere with the assessment of the dysfunctions induced, especially if the dysfunctions are subtle (van den Buuse et al., 2005; van der Staay et al., 2009) and interfere with the normal functioning of the animal, especially in newly created lines (Buehr et al., 2003; Mertens and Rülicke, 2000; Ormandy et al., 2011; van der Staay et al., 2009).

Animal welfare is currently not a scientific but an ethical issue, mainly because of our poor understanding of the concept and the lack of consensus about it. Unfortunately, the concept of animal welfare itself is not well defined and is thus difficult to measure (Ng, 1995; Weerd and Raber, 2005). Animal welfare criteria based on sound scientific evidence are urgently needed to guide the researcher's estimate of the discomfort involved in animal experimentation. In addition, researchers should take societal concerns and society's moral understanding into account (Ohl and van der Staay, 2012) as they play a prominent role in the public's acceptance of animal studies (Buehr et al., 2003). The perceived similarity of the model animal species to humans, the familiarity of the species, and the empathy it evokes strongly affect people's opinion about animal experiments.

A recently introduced concept of animal welfare suggests that animals should have the freedom and capacity to adequately adapt to the demands of prevailing environmental circumstances. This capacity enables the animals to reach a state that they perceive as positive (Ohl and van der Staay, 2012). Positive welfare states are more than the simple exclusion of negative states in animals, such as defined in the "five freedoms." The evaluation whether the demands of environmental conditions can be fulfilled within the limits of an individual's adaptive capacity is as important as investigating whether they induce "negative" states in the animal. Animal welfare as a measure of biological functioning should take the dynamics of an individual's interaction with its environment over time into account.

If the adaptive capacity of genetically modified mice is compromised, it might be necessary to modify husbandry practices. For example, enriched housing improves an animal's welfare. However, it has not yet been investigated whether different types of enrichment differently affect (endo)phenotype expression. A number of studies have shown that the environment (e.g., type of cage) can have large and unanticipated effects on the behavior of genetically engineered mice (Mineur and Crusio, 2009; Oliva et al., 2010) in behavioral tests. On the other hand, recent studies have shown that approaches that incorporate genetic and environmental factors and their interactions more accurately mimic the etiologic factors of neurobehavioral disorders, their underlying pathogenic mechanisms (discussed in Burrows et al., 2011), and consequently improve the model and increase its relevance. Systematic investigation of the effects of environmental conditions (e.g., housing and testing environment; for other potentially relevant factors see Figure 2.2) is mandatory in order to draw conclusions about the effects of experimental manipulations such as genetic modifications and their welfare consequences.

The iterative process of model building and model evaluation

The definition of the purpose(s) of the model is the first step of the model building process (Anisman and Matheson, 2005; van der Staay, 2006; van der Staay et al., 2009) (Figure 2.3). Next, the model is developed and tested. Lastly, the model is validated, taking the questions it is expected to answer, the model's validity (usually face, predictive, and construct validity; Homberg, 2013; Miczek and de Wit, 2008; Sarter and Bruno, 2002) and external validity, i.e., generalizability (Guala, 2003), and animal welfare

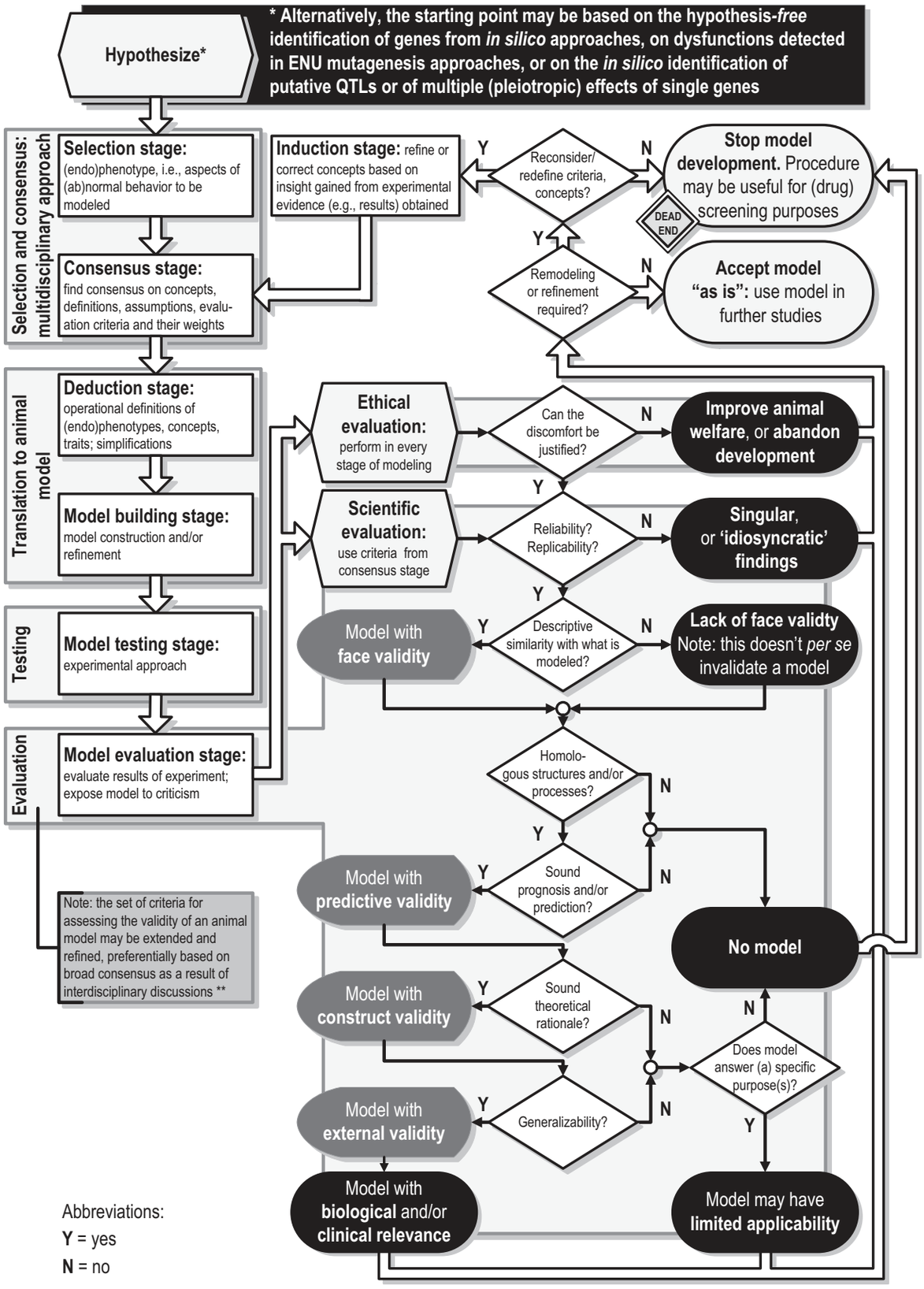

Figure 2.3 Flow diagram depicting model building and model evaluation as an iterative process (inspired by Britt, 1997). Figure modified from van der Staay (2006) with permission from Elsevier.

**: For examples see Belzung and Lemoine (2011). Consensus is urgently needed with respect to the circumstances under which the development of an animal model should be stopped. Such criteria will help to reduce unnecessary use of animals and waste of resources.