

RESEARCH ARTICLE

Validity arguments on the legitimacy and ethics of the forced swim test in rodents

Yingying Han

Institute for Science in Society, Radboud University, the Netherlands
Email: yingying.han@ru.nl

Abstract

Animal models are essential in preclinical research and widely used in drug development, yet their legitimacy has long been debated. These debates intertwine epistemic, pragmatic, social, and ethical considerations. A key criterion for the legitimacy of an animal model is its validity, which assesses how well it serves as a proxy for human disorders and contributes to treatment development. The forced swim test (FST) is a particularly contested case, as criticism regarding its validity has fuelled controversy over its legitimacy. This article examines how validity arguments and non-epistemic factors have been intertwined to shape the legitimacy of the FST as an animal model in depression research. Although public actors have emphasized non-epistemic concerns, including pragmatic, social, and ethical considerations, such as animal welfare, the article shows how they have also utilized the academic controversy about FST's validity to argue against its legitimacy for measuring the efficacy of drugs for human depression.

The paper used the forced swim test and two of my mentors, like the director of my graduate program and our graduate student coordinator, who are both neuroscientists, just started kind of going on and on about how they can't believe this test is still being used, especially it was being used to say whether or not an animal was depressed or in despair, and they were just kind of complaining about how it had been invalidated ... but it keeps being used. So it kind of stuck out to me ... as a really, kind of prime example of how easy it is to get animal studies approved and conduct them even when ... they contribute to poor welfare. You're not really made to justify the science.

Emily Trunnell, a neuroscientist and animal rights activist, now at the forefront of efforts to ban the forced swim test (FST), shared this account with me during an interview when I asked how she first learned about the test.¹ She explained why she chose this 'low-hanging fruit' as a focal point when she began working for People for the Ethical Treatment

¹ Emily Trunnell, online interview with the author, June 2022. I noted that Emily Trunnell played an important role in several campaigns against the FST, and therefore conducted a semi-structured Zoom interview with her to better understand her and PETA's perspective. In addition to the interview, Trunnell kindly provided PETA's emails to multiple pharmaceutical companies and other regulators at my request. These materials, which she agreed could be disclosed, also formed part of the basis for the analysis presented in the section titled 'The validity and legitimacy of the FST'.

of Animals (PETA), an international animal rights organization. Her recollection offers a window into the complexities of the FST controversy, revealing the diverse actors involved – those who use and publish research based on the FST, those who challenge its validity, those who raise ethical concerns, and those in positions to regulate or reform its use. This article delves into these perspectives, analysing different actors' arguments on the validity, legitimacy, and ethics of the FST as a case study of the contestability of animal models.

The conflicting positions surrounding the FST can be better understood when situated within the framework of controversy studies in science and technology studies (STS), which examine how scientific disputes emerge and unfold. Rather than treating controversies as simple disagreements about evidence, this body of work shows how they arise when different actors attach divergent aims, values or expectations to a scientific tool or practice.² The FST debate exemplifies this dynamic: researchers, activists, regulators and industry representatives mobilize distinct concerns when evaluating the test, leading them to challenge or defend its continued use in different ways. From the perspective of controversy studies, such disagreements reveal broader negotiations over which research practices should be maintained, reformed, or abandoned.

The interplay between scientific argument and normative judgement is especially visible in regulatory and legal contexts, and the FST debate reveals similar patterns even outside formal legal arenas. In this article, I adopt an inclusive concept of legitimacy, referring to the socially and institutionally recognized justification for employing a particular animal model in research. Legitimacy arises not only from epistemic adequacy (validity) but also from pragmatic, social and ethical acceptability. In this case, the FST's legitimacy is co-produced by scientific standards, regulatory structures and public moral expectations regarding what an animal model of depression ought to achieve.

Animal models have been considered an indispensable part of preclinical research and are widely used in drug development.³ The practice of using animals in scientific research has a deep historical foundation, evolving from early anatomical studies to their current role as experimental models for understanding human diseases. In the nineteenth century, advances in physiology and experimental medicine cemented the use of animals in laboratories as tools to explore biological processes and disease mechanisms.⁴ Researchers began to recognize the potential of animals to serve as proxies for human systems, driven by a combination of practical availability and their understanding of biological similarities.⁵ By the early twentieth century, animals like rodents were increasingly bred and maintained specifically for research purposes, reflecting the growing importance of creating reliable and reproducible conditions for experimentation.⁶

² Sheila Jasanoff, 'Controversy studies', in George Ritzer and Chris Rojek (eds.), *The Blackwell Encyclopedia of Sociology*, Malden, MA: Wiley, 2019, pp. 1–5.

³ There is a rich and extensive body of historical, science and technology studies (STS), and philosophical scholarship on animal models. This article focuses primarily on how actors use the term, their interpretations of its validity and their critiques, which are analysed with their own definitions. For a broader overview see Rachel Ankeny and Sabina Leonelli, *Model Organisms*, Cambridge: Cambridge University Press, 2020; Sara Green, *Animal Models of Human Disease*, Cambridge: Cambridge University Press, 2024.

⁴ C.A. Logan, 'Before there were standards: the role of test animals in the production of empirical generality in physiology', *Journal of the History of Biology* (2002) 35(2), pp. 329–63.

⁵ R.G.W. Kirk, 'Subjugated love: aligning care with science in the history of laboratory animal research', in Gail Davies, Beth Greenhough, Pru Hobson-West, Robert G.W. Kirk, Alexandra Palmer and Emma Roe (eds.), *What the Humanities and Social Sciences Can Contribute to Laboratory Animal Science and Welfare*, Manchester: Manchester University Press, 2024, pp. 125–51.

⁶ B.T. Clause, 'The Wistar rat as a right choice: establishing mammalian standards and the ideal of a standardized mammal', *Journal of the History of Biology* (1993) 26(2), pp. 329–49.

The development of animals as disease models grew alongside the rise of experimental medicine. In the late nineteenth century, experiments by scientists like Robert Koch used animals to study infectious diseases, demonstrating the causal relationship between specific microbes and illnesses.⁷ These experiments illustrated how animals could be used to study natural biological phenomena, and simulate and manipulate disease states, creating a framework for modern biomedical research.⁸ This shift marked a critical moment in the history of laboratory science, in which animals became essential tools for studying pathology and testing treatments. The integration of animals into laboratory practices transformed scientific inquiry, allowing researchers to investigate complex biological questions that were previously inaccessible.

The twentieth century saw a significant expansion in the use of animal models in psychiatric and neurological research, driven by the rise of behaviourism and the emerging field of psychopharmacology. Early in the century, researchers sought to understand behavioural disturbances in controlled settings, drawing on foundational work by Edward Thorndike and John B. Watson. Thorndike's studies on learning in cats, where he demonstrated the role of trial and error in problem solving, helped establish the idea that behaviour could be studied objectively through measurable responses to stimuli.⁹ Watson later extended this approach, arguing that behaviour, rather than internal mental states, should be the primary focus of psychology.¹⁰ Around the same time, Ivan Pavlov demonstrated that environmental cues could elicit conditioned physiological and behavioural responses, influencing later psychiatric research.¹¹ B.F. Skinner further advanced these ideas by developing operant conditioning techniques using animals like pigeons and rats, showing how reinforcement could shape behaviour over time.¹² This behavioural framework reinforced the role of animal models in studying learning, motivation and maladaptive behaviours, paving the way for later psychiatric research.

The advent of the 'psychopharmacological revolution' in the mid-twentieth century further emphasized the importance of animal models in psychiatry. With the development of drugs like chlorpromazine and imipramine, researchers began using animals to model psychiatric disorders such as depression, anxiety and schizophrenia, creating simplified but insightful representations of human mental illness.¹³ These developments illustrate the historical trajectory of animal models, from tools for studying physiology and infectious diseases to indispensable components of neuroscience and psychiatric research. This historical perspective underscores the pivotal role animal models have played in advancing

⁷ Christoph Gradmann, 'Experimental life and experimental disease: the role of animal experiments in Robert Koch's medical bacteriology', *Futura* (2003) 18(2), pp. 80–8.

⁸ Christoph Gradmann, 'A harmony of illusions: clinical and experimental testing of Robert Koch's tuberculin 1890–1900', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* (2004) 35(3), pp. 465–81.

⁹ E.L. Thorndike, 'Review of animal intelligence: an experimental study of the associative processes in animals', *Psychological Review* (1898) 5(5), pp. 551–3.

¹⁰ J.B. Watson, 'Psychology as the behaviourist views it', *Psychological Review* (1913) 20(2), pp. 158–77.

¹¹ I.P. Pavlov, *Lectures on Conditioned Reflexes: Twenty-Five Years of Objective Study of the Higher Nervous Activity (Behaviour) of Animals*, New York: Liverwright Publishing Corporation, 1928. For more details about Pavlov's life and research see D.P. Todes, *Ivan Pavlov: A Russian Life in Science*, Oxford: Oxford University Press, 2014.

¹² B.F. Skinner, *The Behavior of Organisms: An Experimental Analysis*, Toronto: Appleton-Century, 1938; S. Verhaegh, 'Psychological operationisms at Harvard: Skinner, Boring, and Stevens', *Journal of the History of the Behavioral Sciences* (2021) 57(2), pp. 194–212; A. Rutherford, *Beyond the Box: B.F. Skinner's Technology of Behaviour from Laboratory to Life, 1950s–1970s*, Toronto: University of Toronto Press, 2009.

¹³ Peter M. Haddad, David J. Nutt and A. Richard Green, 'A brief history of psychopharmacology', in Peter M. Haddad and David J. Nutt (eds.), *Seminars in Clinical Psychopharmacology*, Cambridge: Cambridge University Press, 2020, pp. 1–34.

our understanding of complex biological and psychological phenomena. In the context of depression, animal models continue to provide critical insights, offering a controlled means to investigate its underlying mechanisms and test novel treatments.¹⁴

At the same time, the legitimacy of using animal models has been debated almost from the start, with epistemic, pragmatic, social, and ethical considerations often intertwined.¹⁵ The ethical concerns about animal research have a long history. For example, the antivivisection movement, which gained particular prominence in Britain in the late nineteenth century, brought public attention to the moral implications of using animals in laboratories.¹⁶ Campaigns against vivisection questioned the justification of experiments that caused pain and suffering, leading to early legislative efforts for the protection of animal rights. These historical debates laid the groundwork for modern approaches to animal welfare, introducing moral considerations that continue to shape the discourse around animal research today.

Another key aspect of these debates centres on the validity of animal models, which focuses primarily on epistemic considerations but indirectly plays an important role in the cost-benefit analysis within ethics debates. Technically, the validity of animal models concerns the extent to which the model serves as a plausible proxy for understanding a human disorder and thus contributes to the development of potential treatments.¹⁷ Among the many validity frameworks proposed for animal models, the most widely adopted one was developed by clinical psychologist Paul Willner, renowned for his work on animal models of depression.¹⁸ His framework has been used to evaluate animal models in academic literature and public debates surrounding the legitimacy and ethics of animal research, which will be discussed further in following sections.

Arguments about validity and legitimacy have been utilized in addition to ethical considerations in animal rights campaigns, closely following trends in academic discussions, but with their own twist. One case this paper will explore is the academic and public debate around the legitimacy of a specific animal model of depression: the FST. Since 2018, an international animal rights organization, People for the Ethical Treatment of Animals (PETA), has questioned the validity and ethical justification of the FST in research and drug development. This advocacy has led several pharmaceutical companies to ban the test while also prompting reactions from universities, journals and regulatory authorities. By analysing the role of validity arguments in these debates, this article explores how epistemic and ethical considerations intersect in shaping the legitimacy of animal models in science.

This article will show how arguments about the validity of the FST have been used in the controversy about its legitimacy. I aim to demonstrate that, in this case, the epistemic issue of the FST's validity for a range of purposes, intertwining with non-epistemic values including pragmatic and ethical concerns, plays an essential role in determining its legitimacy in both academic and public discussions. I will introduce the FST and outline its

¹⁴ L. Gerber, *Le laboratoire des esprits animaux: Modéliser le trouble mental à l'ère de la psychopharmacologie*, Lausanne: Editions BHMS, 2022.

¹⁵ C. Shelley, 'Why test animals to treat humans? On the validity of animal models', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* (2010) 41(3), pp. 292–9.

¹⁶ A.W.H. Bates, *Anti-vivisection and the Profession of Medicine in Britain*, London: Palgrave Macmillan UK, 2017.

¹⁷ Y. Han, 'Multiple historic trajectories generate multiplicity in the concept of validity', *Perspectives on Science* (2024) 32(4), pp. 488–517.

¹⁸ Paul Willner has a background in psychology and physiology, graduating from Oxford, where he got his PhD in 1974. He held psychology chairs at London Metropolitan University and Swansea University and was a fellow of the British Psychology Society. Late in his career, he trained as a clinical psychologist. He was also active in pharmacology, with a presidency of the European Behavioural Pharmacology Society (2000–1), and he is one of the founding editors of the journal *Behavioural Pharmacology*.

historical development in the next section. Then follows a discussion of how researchers have assessed the FST's validity in academia, especially using Willner's classic theoretical framework.¹⁹ The final section before the conclusion traces the FST legitimacy controversy beyond academia, focusing on validity arguments used by non-academic actors such as animal rights organizations and regulators.

History of the FST

Development of the FST

Around 1960, the first compounds from the two primary classes of early antidepressant, tricyclic antidepressants (TCAs) and monoamine oxidase inhibitors (MAOIs), were developed and introduced in Europe and the US, resulting in significant interest in new antidepressant development and testing.²⁰ In 1977, Roger D. Porsolt's research group developed the forced swim test (FST) as a behavioural animal model to screen antidepressants.²¹ Porsolt was then affiliated with the research and development departments of two pharmaceutical companies, Synthélabo and Centre de recherche Delalande, which partly explains his team's focus on demonstrating the FST's validity in testing new antidepressants, including TCAs and MAOIs.²²

The standard procedure of the FST starts with a rodent (typically a rat or a mouse) placed in an inescapable cylinder with water. The animal initially swims until it becomes immobile or starts floating (defined as exhibiting only movements necessary to keep the nose above the water surface; see Figure 1). The main behavioural outcome is measured by the latency to and duration of immobility.²³

According to Porsolt's recollection, the immobility phenomenon was first observed in a water maze experiment: 'Intuitively it seemed that for these rats the situation was insoluble and that they had simply given up hope ("behavioral despair")'.²⁴ The FST was thus originally often referred to as the 'behavioural despair' test.²⁵ Although Porsolt and colleagues did not explicitly cite it, their use of the term 'despair' may have been influenced by contemporary primate research on separation by Harry Harlow and colleagues. These studies documented behavioural patterns in separated infant monkeys that aligned with the protest–despair sequence identified in earlier human separation research, and the work was widely discussed in the 1960s and 1970s.²⁶ It is plausible that Porsolt's group adopted similar terminology either to draw on this visibility or to suggest an implicit theoretical continuity between rodent and primate models.

¹⁹ Paul Willner, 'Validation criteria for animal models of human mental disorders: learned helplessness as a paradigm case', *Progress in Neuropsychopharmacology and Biological Psychiatry* (1986) 10(6), pp. 677–90.

²⁰ Haddad, Nutt and Green, op. cit. (13).

²¹ R.D. Porsolt, M. Le Pichon and M. Jalfre, 'Depression: a new animal model sensitive to antidepressant treatments', *Nature* (1977) 266(5604), pp. 730–2.

²² In 1985 Porsolt founded a preclinical Contract Research Organization (CRO), now known as Porsolt SAS, and remained president until 2014.

²³ R.D. Porsolt, G. Anton, N. Blavet and M. Jalfre, 'Behavioural despair in rats: a new model sensitive to antidepressant treatments', *European Journal of Pharmacology* (1978) 47(4), pp. 379–91.

²⁴ R.D. Porsolt, 'Behavioral despair: past and future', in Bernard Lerer and Samuel Gershon (eds.), *New Directions in Affective Disorders*, New York: Springer, 1989, pp. 17–20, 17; J.A. Sullivan, 'Reconsidering "spatial memory" and the Morris water maze', *Synthèse* (2010) 177(2), pp. 261–83.

²⁵ Porsolt op. cit. (24).

²⁶ S.J. Suomi, H.F. Harlow and C.J. Domek, 'Effect of repetitive infant–infant separation of young monkeys', *Journal of Abnormal Psychology* (1970) 76(2), pp. 161–72.



Figure 1. Rat immobility state in the FST. R.D. Porsolt, G. Anton, N. Blavet and M. Jalfre, 'Behavioural despair in rats: a new model sensitive to antidepressant treatments', *European Journal of Pharmacology* (1978) 47(4), pp. 379–91, Figure 1.

Porsolt and colleagues claimed that their 'test procedure reproduces some aspects of human depression' in animals.²⁷ They argued that their interpretation of immobility as 'a state of lowered mood or hopelessness' was supported by various experiments involving pharmacological and other treatments considered effective in depression.²⁸ The behavioural immobility in both the rats and mice tests was found to be decreased by a wide range of antidepressants injected before the test period. Thus the FST was proposed as 'a new model sensitive to antidepressant treatments' and a method 'capable of discovering new types of antidepressant agents hitherto undetectable using classical screening tests', focusing mainly on its application in drug development.²⁹

²⁷ Porsolt *et al.*, op. cit. (23), p. 388.

²⁸ Porsolt *et al.*, op. cit. (23), p. 386.

²⁹ Porsolt, Le Pichon and Jalfre, op. cit. (21), p. 732.

Although Porsolt's stated aim was to develop a rapid screening tool for antidepressant compounds, their model's name choice using 'despair' and their attempt to link the model to the phenomenon of 'learned helplessness' gave the test a dual identity from the outset. This ambiguity as a partially pragmatic screening tool and partially putative model of depression may have helped facilitate the later broadening of its application and also planted the seeds of recent controversy, as different communities highlighted different aspects of the test and applied distinct criteria when assessing its validity.

Challenges and modifications of the FST

Not long after the FST was developed, its interpretation was questioned. One line of criticism targeted the interpretation of immobility as 'behavioural despair'. For instance, Hawkins and colleagues replicated and analysed the animals' behaviours in the FST and found evidence suggesting the animals were less stressed in test sessions showing higher immobility.³⁰ Hence they argued that immobility was 'an adaptive response to a stressful situation'.³¹ Although supported by many other studies, this interpretation 'did not have a great impact on the way researchers have later interpreted the test'.³² According to the analysis by Marc L. Molendijk and Edo Ronald de Kloet, the majority of publications reporting FST results interpreted immobility as depressive-like behaviour (72 per cent in 2014–18, 58 per cent in 2018–20) or inferred responses to substances with potential antidepressant properties (19 per cent in 2014–18, 30 per cent in 2018–20).³³ References to other interpretations (e.g. adaptive coping) or uses were the minority (10 per cent in 2014–18, 14 per cent in 2018–20).³⁴

Apart from the theoretical concerns, researchers also found that various factors could influence the FST outcome. The list of these factors has grown since 1977, including parameters in the experimental set-up, procedures before the test, and characteristics of the animals, such as sex, age and strain.³⁵ Some of the factors were acknowledged and led to FST protocol modifications. For instance, using deeper water allowed researchers to characterize and quantify different types of active behaviours, such as climbing, swimming and diving, in addition to immobility, which enabled the FST to provide a richer and more nuanced output profile.³⁶ With this modification, the FST could account for some previous negative results with a major type of antidepressant (i.e. selective serotonin re-uptake inhibitors – SSRIs), further establishing its use to screen antidepressants.³⁷ As presented

³⁰ J. Hawkins, R.A. Hicks, N. Phillips and J.D. Moore, 'Swimming rats and human depression', *Nature* (1978) 274(5670), p. 512.

³¹ Hawkins *et al.*, op. cit. (30), p. 512.

³² A. Armario, 'The forced swim test: historical, conceptual and methodological considerations and its relationship with individual behavioral traits', *Neuroscience and Biobehavioral Reviews* (2021) 128, pp. 74–86, 76.

³³ M.L. Molendijk and E.R. de Kloet, 'Coping with the forced swim stressor: current state-of-the-art', *Behavioural Brain Research* (2019) 364, pp. 1–10.

³⁴ M.L. Molendijk and E.R. de Kloet, 'Forced swim stressor: trends in usage and mechanistic consideration', *European Journal of Neuroscience* (2022) 55(9–10), pp. 2813–31.

³⁵ O. Bogdanova, S. Kanekar, K.E. D'Anci and P.F. Renshaw, 'Factors influencing behavior in the forced swim test', *Physiology and Behavior* (2013) 118, pp. 227–39.

³⁶ M.J. Detke, M. Rickels and I. Lucki, 'Active behaviors in the rat forced swimming test differentially produced by serotonergic and noradrenergic antidepressants', *Psychopharmacology* (1995) 121(1), pp. 66–72.

³⁷ While reactivity to SSRIs was important in the process of validation in the 1980s and 1990s, they lost this value in more recent times, because now research focuses more on non-responders to SSRIs. See C. Belzung, 'Innovative drugs to treat depression: did animal models fail to be predictive or did clinical trials fail to detect effects?', *Neuropsychopharmacology* (2014) 39(5), pp. 1041–51; L. Keuck, 'Scope validity in medicine', in M. Schermer and N. Binney (eds), *A Pragmatic Approach to Conceptualization of Health and Disease*, Cham: Springer, 2024, pp. 115–33.

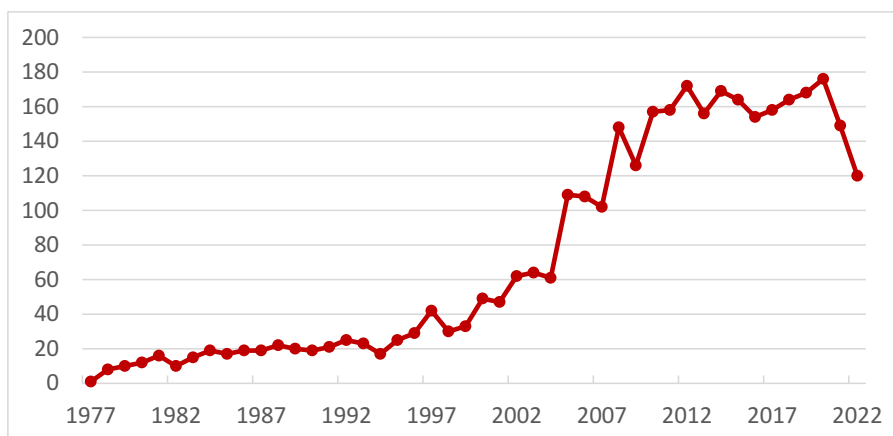


Figure 2. Web of Science citation counts of the original paper introducing the FST: R.D. Porsolt, A. Bertin and M. Jalfre, 'Behavioral despair in mice: a primary screening test for antidepressants', *Archives internationales de pharmacodynamie et de thérapie* (1977) 229(2), pp. 327–36.

in a 2013 overview of factors of influence, the modification was seen as a common calibration/optimization endeavour that did not undermine the legitimacy of the FST: 'The FST, in its classical or modified version, provides a unique opportunity to assess antidepressant efficacy in a rapid, low-cost, and reliable manner ... the FST should be standardized and [the factors that may induce variation in immobility] should be considered during the study design.'³⁸

The uptake of the FST

Although the FST has become 'the most widely used tool for assessing antidepressant activity preclinically', it was not so popular before the 2000s.³⁹ Its 'popularity grew in the early 2000s when scientists began modifying mouse genomes to mimic mutations linked to depression in people'.⁴⁰ Transgenic models of depression served as powerful tools for exploring novel molecular targets for antidepressant treatments, particularly when conventional pharmacological approaches were insufficient. Behavioural changes in these genetically modified mice were often assessed using standardized tests such as the FST. Figure 2 illustrates the Web of Science citation counts of the publication that first introduced the FST, showing the increased uptake of the FST since around 2004. In the 'genomic era', researchers published articles to discuss the FST's utility for studying depression-related behaviour in genetically modified animals.⁴¹ For instance, Porsolt evaluated several animal models of depression 'for their suitability for transgenic research' and suggested that the 'great advantage of [the FST] is its procedural simplicity and its reproducibility', stressing pragmatic reasons to promote the FST in genetic research.⁴²

³⁸ Bogdanova *et al.*, *op. cit.* (35), pp. 234–5.

³⁹ J.F. Cryan, 'Depression', in George F. Koob, Michel Le Moal and Richard F. Thompson (eds.), *Encyclopedia of Behavioral Neuroscience*, Boston, MA: Elsevier, 2010, pp. 382–6.

⁴⁰ S. Reardon, 'Depression researchers rethink popular mouse swim tests', *Nature* (2019) 571(7766), pp. 456–7.

⁴¹ A.E. Guttmacher and F.S. Collins, 'Welcome to the genomic era', *New England Journal of Medicine* (2003) 349(10), pp. 996–8.

⁴² R.D. Porsolt, 'Animal models of depression: utility for transgenic research', *Reviews in the Neurosciences* (2000) 11(1), pp. 53–8, 54.

Cryan and Mombereau listed forty genetically modified mouse strains with depressive phenotypes, thirty-five of which were assessed using the FST.⁴³ The authors speculated that the FST's predominance was due to 'its relative reliability across laboratories and its ability to detect activity in a broad spectrum of clinically effective antidepressants'.⁴⁴ This indicates that the FST may have become part of the standards by which transgenic models of depression were assessed by the early 2000s. As the FST developed into one of the standard behavioural assays for evaluating transgenic lines, it contributed to defining which phenotypes counted as 'depressive-like'. In this way, it has influenced the emerging molecular-genetic narrative of depression by privileging pathways associated with drugs that reduced immobility, and it may eventually subtly shape our understanding of the genetic basis of depression. With its increasing popularity, the FST has extended its application scope from its original development as an antidepressant discovery test to an animal model of depression that reflected 'depressive-like behaviours' in rodents and beyond.

The recent controversy

In the past decade, criticisms have been raised regarding the FST's use as an animal model of depression. For instance, Nestler and Hyman stated that the FST was 'not [a] model of depression at all' and it was a 'black box test' that involved an 'enormous anthropomorphic leap'.⁴⁵ Molendijk and De Kloet also published multiple articles arguing that immobility in the FST 'does not reflect depression'. They rekindled the 1980s and 1990s challenges to the FST, especially supporting the alternative interpretation of immobility as an adaptive coping behaviour.

Despite these criticisms in the past decade, the FST remained popular in academia. Molendijk and De Kloet calculated the annual number of publications in PUBMED reporting FST results and estimated that one FST paper per day was published from 2015 to 2020.⁴⁶ They noted that 'the FST for antidepressant drug screening was further diminished, but that its application for phenotyping of animal behaviour remained popular and was extended in recent years to test interventions in stressed-out animals'.⁴⁷ However, the trend seems to have changed since 2022. Using their search string, I found a decrease in the number of publications in PUBMED mentioning FST (574 times in 2022 versus 622 times in 2021 and 653 times in 2020), which was consistent with the decreased citation counts of the original FST paper from Web of Science (see Figure 2). The FST's seemingly diminishing popularity may be an indication of increased concerns about its legitimacy.

The validity of the FST

To assess the validity of the FST, Paul Willner's criteria are most frequently used. His validity framework was built partially upon McKinney and Bunney's introduction of 'minimal requirements' for an animal model of depression and partially by adapting psychometric

⁴³ J.F. Cryan and C. Mombereau, 'In search of a depressed mouse: utility of models for studying depression-related behavior in genetically modified mice', *Molecular Psychiatry* (2004) 9(4), pp. 326–57.

⁴⁴ Cryan and Mombereau, op. cit. (43), p. 340.

⁴⁵ E.J. Nestler and S.E. Hyman, 'Animal models of neuropsychiatric disorders', *Nature Neuroscience* (2010) 13(10), pp. 1161–9, 1166.

⁴⁶ M.L. Molendijk and E.R. de Kloet, 'Immobility in the forced swim test is adaptive and does not reflect depression', *Psychoneuroendocrinology* (2015) 62, pp. 389–91; Molendijk and De Kloet, op. cit. (34).

⁴⁷ Molendijk and De Kloet, op. cit. (34), p. 6.

notions of validity for preclinical research.⁴⁸ Willner's classic criteria feature face, predictive and construct validity. When applied to animal models of depression, face validity was considered guaranteed by the model's symptomatic resemblance to depression, which includes similarity in both behavioural symptoms and responses to treatments. As for predictive validity, whether the model could correctly identify various antidepressants, its false-positive/-negative prediction rates, and the correlation between antidepressant potency in the model and in clinical trials became important indicators. While establishing the face and predictive validity for animal models of depression relies heavily on antidepressants and other treatment effects, construct validity refers more to the link between the interpretation of the behaviours in the model and the empirical and theoretical grounds for the features being modelled.

Using these criteria, Willner assessed the FST multiple times, such as in his 1984, 1986, 1990 and 2006 articles about validity criteria for animal models of depression.⁴⁹ In all these articles, he questioned the face and construct validity of the FST. One major criticism of face validity focused on the different temporal characteristics of antidepressant responses in the animal model versus in human patients. Another major concern was that the analogy between the FST and depression relied solely on the subjects' inability or reluctance to maintain effort, which was connected to construct validity concerning the underlying theoretical rationale. No explicit theoretical rationale was proposed for its status as an animal model of depression since it was originally developed merely as 'a new model sensitive to antidepressant treatments'.⁵⁰ The recent criticisms of the FST's construct validity mainly focused on the alternative interpretation of immobility as adaptive coping (e.g. conserving energy) instead of 'a state of despair' (i.e. 'giving up trying'), which was rarely mentioned by Willner. According to Willner, the FST's construct validity was derived from an unjustified link to another animal model of depression (i.e. the 'learned-helplessness' model).⁵¹

Despite its questionable face and construct validity, the FST was known for its high predictive validity, particularly for identifying antidepressants, which Willner also acknowledged: 'There is, in fact, a significant correlation between the potency of antidepressants in the 'behavioural despair' test [i.e. the FST] and their clinical potency, which has not been demonstrated for any other animal model of depression.'⁵² Various treatments effective in human patients (including most antidepressants, electroconvulsive therapy and REM sleep deprivation) also reduced immobility or delayed its onset, showing that the FST could predict effective treatments in clinical settings and thus had high predictive validity.⁵³ Although there were reports of false positives and false negatives and recent reviews that questioned the FST's ability to predict the dose response of antidepressants in human patients or its utility for developing novel antidepressants, the consensus remained that the FST was highly predictive for a wide range of depression treatments.⁵⁴ Even reviews

⁴⁸ Walter T. McKinney and William E. Bunney, 'Animal model of depression', *Archives of General Psychiatry* (1969) 21(2), pp. 240–8, 240. See also Willner, op. cit. (19); Han, op. cit. (17), for a detailed discussion.

⁴⁹ Willner, op. cit. (19); P. Willner, 'The validity of animal models of depression', *Psychopharmacology* (1984) 83(1), pp. 1–16; Willner, 'Animal models of depression: an overview', *Pharmacology & Therapeutics* (1990) 45(3), pp. 425–55; P. Willner and P.J. Mitchell, 'Animal models of depression', in M. Koch (ed.), *Animal Models of Neuropsychiatric Diseases*, London: Imperial College Press, 2006, pp. 223–92.

⁵⁰ Porsolt, Le Pichon and Jalfre, op. cit. (21), p. 730.

⁵¹ Willner, 'The validity of animal models of depression', op. cit. (49).

⁵² Willner, 'Animal models of depression', op. cit. (49), p. 437.

⁵³ Willner and Mitchell, 'Animal models of depression', op. cit. (49).

⁵⁴ N.Z. Kara, Y. Stukalin and H. Einat, 'Revisiting the validity of the mouse forced swim test: systematic review and meta-analysis of the effects of prototypic antidepressants', *Neuroscience & Biobehavioral Reviews* (2018)

critical of the FST as a valid animal model of depression tended to acknowledge its high predictive validity, which contributes to its immense popularity.⁵⁵

Considering, on the one hand, the relatively low face and construct validity and, on the other hand, the FST's high predictive validity, assessing its overall validity is not trivial. One way to determine which validity criteria to apply would be to consider *for what purpose* the FST is used, or how its results are interpreted. So far, this section has mainly discussed the validity criteria for animal models of depression. However, these criteria have also been used to assess the FST for other purposes, including as an antidepressant screening tool. Willner and Mitchell noted that different validity criteria were required for each purpose:

These different functions of animal models have different, and to some extent, conflicting requirements. A simulation of depression aims to mimic aspects of the clinical situation, and should embody a degree of complexity, to permit investigation of the validity of the model, as well as a slow onset of antidepressant action, comparable to the clinical time course. By contrast, the only essential requirement for antidepressant screening tests is that they make accurate predictions of antidepressant activity. For practical reasons, they should also be cheap, robust, reliable and easy to use and for all of these subsidiary reasons, a screening test should in principle be as simple as possible.⁵⁶

In addition to epistemic considerations such as various validity criteria, pragmatic factors also come into play in practice. However, Molendijk and De Kloet observed that validity issues were not discussed, or were discussed improperly.⁵⁷ Nestler and Hyman argued that a 'frequent failing of the literature is the use of such screens as if they were based on validated pathophysiological models'.⁵⁸ Sometimes pragmatic reasons, such as the FST's low requirement for equipment, training and execution, were considered, but they are rarely the highlight of the discussion.⁵⁹ In a review that analysed criteria to select model organisms, Michael Dietrich and co-authors compiled a list of twenty criteria (see Table 1). Many of them point to practical and social considerations in addition to the epistemic ones.⁶⁰

Underlining the importance of test use in validity, some researchers started to argue for the FST's validity *for another purpose*, namely to study stress and coping instead of depression.⁶¹ At first glance, their scathing criticism of the FST's validity might lead to the false conclusion that they questioned its legitimacy. However, these academic critics did not advocate for its discontinuation. Rather, they tried to repackage the FST as 'a powerful paradigm to examine the mechanism underlying coping with inescapable stressors'.⁶² One

84, pp. 1–11; E.R. Trunnell and C. Carvalho, 'The forced swim test has poor accuracy for identifying novel antidepressants', *Drug Discovery Today* (2021) 26(12), pp. 2898–904.

⁵⁵ Nestler and Hyman, op. cit. (45); Armario, op. cit. (32); Molendijk and De Kloet, op. cit. (33); Molendijk and De Kloet, op. cit. (46); E.R. de Kloet and M.L. Molendijk, 'Coping with the forced swim stressor: towards understanding an adaptive mechanism' *Neural Plasticity*, 2016, article 6503162, at <http://dx.doi.org/10.1155/2016/6503162>.

⁵⁶ Willner, 'Animal models of depression', op. cit. (49), pp. 223–4.

⁵⁷ Molendijk and De Kloet, op. cit. (46).

⁵⁸ Nestler and Hyman, op. cit. (45), p. 1163.

⁵⁹ Porsolt, op. cit. (42).

⁶⁰ M.R. Dietrich, R.A. Ankeny, N. Crowe, S. Green and S. Leonelli, 'How to choose your research organism', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* (2020) 80, 101227. Model organisms are not the same as 'animal models', as we loosely use the term in this article, but there are overlaps. For a detailed delineation see note 3 above.

⁶¹ Armario, op. cit. (32); Molendijk and De Kloet, op. cit. (33); Molendijk and De Kloet, op. cit. (46).

⁶² Molendijk and De Kloet, op. cit. (33), p. 7.

Table 1. One example of criteria to select model organisms. M.R. Dietrich, R.A. Ankeny, N. Crowe, S. Green and S. Leonelli, ‘How to choose your research organism’, *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* (2020) 80, 101227, Table 1.

Cluster	Criteria
(A) Access	(1) Ease of supply
	(2) Phenomenal access
	(3) Ethical considerations
(B) Tractability	(4) Standardization
	(5) Viability and durability
	(6) Responsiveness
	(7) Availability of methods and techniques
	(8) Researcher risks
(C) Resourcing	(9) Previous use
	(10) Epistemic resources
	(11) Training requirements
	(12) Informational resources
(D) Economies	(13) Institutional support
	(14) Financial considerations
	(15) Community support
	(16) Affective and cultural attributes
(E) Promise	(17) Commercial and other applications
	(18) Comparative potential
	(19) Translational potential
	(20) Novelty

major issue raised by critics like Molendijk and De Kloet was that the FST does not model despair or chronic stress in depression, but rather an active coping mechanism for acute stress. Consequently, they proposed using the FST to study the neural mechanisms of coping and acute stress rather than depression.⁶³ Commons and colleagues, who support this perspective, suggest employing the FST to investigate maladaptive acute stress responses in autism spectrum disorder.⁶⁴ While these alternative interpretations challenge the FST’s validity for its original purpose, they also present new opportunities for repurposing the test. Thus the FST’s legitimacy could be retained by shifting the purpose for which it is valid.

⁶³ Molendijk and De Kloet, op. cit. (34).
⁶⁴ K.G. Commons, A.B. Cholanians, J.A. Babb and D.G. Ehlinger, ‘The rodent forced swim test measures stress-coping strategy, not depression-like behavior’, *ACS Chemical Neuroscience* (2017) 8(5), pp. 955–60.

The validity and legitimacy of the FST

Validity arguments used by animal rights organizations

Beyond academic communities, the controversy about the FST's validity went public in 2018, when People for the Ethical Treatment of Animals (PETA) started campaigns to urge big pharmaceutical companies to ban the FST. Between 1988 and 2017, an average of almost nine English-language articles about the FST appeared in the global media each year, mostly in a neutral tone or positive, such as related to drug development. The only negative article appeared in 2013, written by the animal rights organization Animal Aid. However, after PETA's November 2018 report, the rate went up to 334 articles in four years (2018–22), with the majority negative about the test, relating to the animal rights campaign.⁶⁵ In most of the news entries, the animal rights advocates invoked ethical concerns about the FST, calling it a 'cruel' 'near-drowning test'.

In contrast to the often neutral or even positive attitude toward using animal models in the scientific literature, the stance of animal rights activists was against animal experimentation in general.⁶⁶ In the case of the FST, PETA not only appealed to ethical concerns to undermine its legitimacy, but also drew on scientific discussions about the FST and emphasized that the test is 'scientifically worthless'.⁶⁷ In the emails PETA sent to various pharmaceutical companies (e.g. Eli Lilly and Company, Bristol-Myers Squibb, AbbVie, and Pfizer), they pointed out that 'the applicability of an animal's behaviour during the FST to their mood, or to human depression, or to the utility of a compound for treating human depression has been refuted'.⁶⁸ This was followed by 'a thorough discussion of this matter' in an attached document entitled 'The invalidity of the forced swim test' written by Emily Trunnell, a neuroscience PhD, who is the main advocate against the FST in PETA-US.

In this document, Trunnell summarized the development of the FST and presented the scientific controversy around it, emphasizing the alternative interpretation of immobility in the FST as a learned adaptive behaviour. Many scientific articles were cited, including discussions about the lack of face and construct validity in the FST.⁶⁹ Modifications of the FST were noted by Trunnell as evidence that 'further invalidate[s] the FST as a reliable measure of "despair" or behavior in general'. She concluded, 'it's problematic that in order to make this test "work", experimenters make extensive modifications in order to get the results they want in each situation – for different species or different drugs'.⁷⁰ The modifications that were considered improvements in the scientific literature were therefore seen as signs of the FST's weakness by PETA.

While PETA drew heavily on the abundant academic findings to argue against FST's construct validity and face validity, Trunnell also presented PETA's own analyses to undermine the predictive validity of the FST. In the six-page document, at least four pages were devoted to construct validity and face validity, whereas only two damning sentences were spent on predictive validity:

⁶⁵ I searched for news articles in Nexis Uni in March 2022 with the search string 'forced NEAR/1 swim NEAR/10 test', finding 584 hits.

⁶⁶ On PETA's official website, an article entitled 'Animals used for experimentation' illustrates their opinions, in which they reported that their 'dedicated team of scientists and other staff members work full time exposing the cruelty of animal tests in order to ensure their imminent end'. See www.peta.org/issues/animals-used-for-experimentation (accessed 6 February 2023).

⁶⁷ Information from People for the Ethical Treatment of Animals (PETA), 'Pharmaceutical behemoth ends near-drowning tests on animals' (2018), at www.peta.org/media/news-releases/pharmaceutical-behemoth-ends-near-drowning-tests-on-animals.

⁶⁸ Trunnell, op. cit. (1).

⁶⁹ For example, see Commons *et al.*, op. cit. (64); Molendijk and De Kloet, op. cit. (46).

⁷⁰ Trunnell, op. cit. (1).

In December 2018, PETA scientists began analyzing publicly available data on pharmaceutical companies' use of the FST to test novel compounds for their potential value as human antidepressants. PETA found that the predictive efficacy of the FST in these cases was less than 50 percent.⁷¹

Providing more solid evidence, Trunnell and Carvalho, the latter not associated with PETA, published a review in a scientific journal, focusing on the FST's lack of predictive validity as an antidepressant screening tool.⁷² Instead of referring to the validity criteria, they mainly use terms like accuracy, specificity and sensitivity, which can be considered important components of predictive validity and are more commonly used in pharmacological contexts (as explained in an interview with Trunnell). This article concluded,

FST is not a reliable tool for the screening of antidepressant drugs at the pharmaceutical level ... However, the FST is acutely stressful to animals; thus, the practices and, in some countries, the law (e.g., European Union), requires that it is only conducted after a cost-benefit analysis, that is, when the expected benefits (understanding and treating human depressive disorders) would outweigh the costs (harm to the animals).⁷³

This line of reasoning was also prominent in my interview with her, in which she pointed out that the scientific validity of the FST plays a crucial role in assessing the benefit of its use. When arguing against the legitimacy of the FST, PETA not only appealed to ethical concerns (the 'harm') but also persistently argued against its validity and reliability (the 'benefit'). Considering the low epistemic value and high ethical cost, PETA concluded that the FST should be banned.

While the academic controversy did not significantly affect FST use in publications, the PETA campaign seemed more consequential. Since 2018, PETA has claimed to have convinced thirteen big pharmaceutical companies to discontinue using the FST. A few universities also no longer approve studies using the FST after communications with PETA. When asked what arguments and/or actions were more influential or decisive in these successes, Emily Trunnell replied in the interview,

It is hard to tell sometimes, honestly, if they [i.e. the companies or universities] are responding to ... 'we just don't want negative press from PETA so we'll agree' or if they are responding to the actual ... like looking at the science. I do know in some cases their responses have been ... you know ... 'we agree that it's not a useful test'.

This suggests that, in addition to epistemic reasoning, a combination of non-epistemic values affected the legitimacy of the FST in the public domain and in regulating its use for drug development and research in an informal way (as opposed to formal legislation).

Interaction between animal rights organizations and governmental regulators

PETA's efforts to approach journal editors and regulatory actors such as the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have so far led to limited direct responses and no formal restrictions regarding the use of the FST. While PETA

⁷¹ Trunnell, op. cit. (1).

⁷² Trunnell and Carvalho, op. cit. (54).

⁷³ Trunnell and Carvalho, op. cit. (54), p. 2901.

has been the main driving force in animal rights movements against the legitimacy of using the FST for research and pharmaceutical purposes, other animal rights organizations were also actively engaged. One interesting example is a petition by two New Zealand animal rights organizations aiming to '[e]nd the use of the forced swim test in New Zealand'. The petition was submitted to the Economic Development, Science and Innovation Committee, the regulator for animal research in New Zealand. This case was well documented, and the materials are publicly available, presenting not only the animal rights arguments but also the direct response from the regulator, making it a nice window on often obscured regulatory deliberations.⁷⁴

In October 2019, on behalf of the local animal rights organizations the New Zealand Anti-Vivisection Society (NZAVS), SAFE and others who signed online, NZAVS campaigner Tara Jackson submitted a petition requesting 'the Economic Development, Science and Innovation Select Committee to take immediate action to pass legislation to immediately ban the Forced Swim Test and to conduct a formal review and evaluation of the validity of all animal-based psychological tests used in New Zealand'. A twenty-one-page document was submitted with two appendices, including a version of Emily Trunnell's article 'The invalidity of the forced swim test'. In the main text, the petitioners outlined six arguments for the ban, putting 'The invalidity of the forced swim test' at the top of the list, suggesting its prominent role in arguing against its legitimacy. In the executive summary directly following the outline of arguments, the invalidity of the FST was summarized as 'The Forced Swim Test does not reliably predict the human response – nullifying any scientific justification for carrying out the test'. This argument focuses solely on predictive validity and implies that predictability is necessary to justify the use of the FST. Other validity criteria were not explicitly mentioned, and a detailed discussion was directed to the appendix by Trunnell (similar to the article discussed in the previous section).

A subsection was titled with a rhetorical question: 'If the Forced Swim Test is so flawed, why is it still being used?' The answer appealed mainly to the continuation of a certain research tradition:

The claim is often made that the Forced Swim Test is commonly used to model human [d]epression. This claim appears to be the strongest justification that researchers have for why they still use the test today. As a result, the Forced Swim Test is used 'widely' without sufficient evidence validating its translational value to humans.

However, in the scientific literature, social factors such as scientific inertia, as suggested by this quote, were seldom mentioned when arguing for the validity and legitimacy of the FST.⁷⁵

Apart from the written material, oral evidence was also presented to the committee in a recorded meeting.⁷⁶ Three animal rights representatives, including NZAVS's Tara Jackson, were physically present, and PETA's Emily Trunnell joined via Zoom. Craig Johnson, professor of veterinary neurophysiology and animal welfare science, and director of research ethics at Massey University, was invited by the committee to serve as the scientific and

⁷⁴ 'Petition of Tara Jackson on behalf of the NZ Anti-Vivisection Society, SAFE, and 7,861 others: end the use of the Forced Swim Test in New Zealand', 7 October 2019, at www.parliament.nz/en/pb/petitions/document/PET_91564/petition-of-tara-jackson-on-behalf-of-the-nz-anti-vivisection?fbclid=IwAR2z3fzq65Ep0Do4vCjd7-MMJyVHdPNHSy_GucJ3y3XcZBXN9mqskL2IHTc (accessed 6 February 2023).

⁷⁵ S. Lohse, 'Scientific inertia in animal-based research in biomedicine', *Studies in History and Philosophy of Science* (2021) 89, pp. 41–51.

⁷⁶ Economic Development Science and Innovation Committee Official Facebook Account, 'Petition of Tara Jackson (20 February 2020)', at www.facebook.com/EDSISCNZ/videos/405422310308020 (accessed 6 February 2023).

ethics expert. While the animal rights advocates' oral evidence did not add much to that in the written material, Johnson's oral evidence turned out to be quite influential in the committee's final decision.

As to the validity of the FST, Johnson's opinion was in line with the academic discussion (as discussed in the section above titled 'The validity of the FST'), and he proposed that the best way to change the use of the FST would be through education for ethics committees. He concluded,

I would be ... hmm ... opposed to banning an individual test, partly because of the difficulty in pinning down the definition of the test ... say what it was that we're banning but also because it would complicate regulatory space and I think that will detract the clarity that we have at the moment.⁷⁷

In the final report from the regulator, the petitioners' validity arguments were recognized: 'The petitioners believe that the FST holds little to no validity as a measure of depression or human antidepressant efficacy and does not benefit drug development.' The official response to the petition echoes the expert's (i.e. Johnson's) opinion:

We do not believe legislation is necessary to end the use of the Forced Swim Test. The test is used infrequently in New Zealand, and we heard that its use in academic studies is not likely to continue into the future.

We support the continuing education of the ethics boards of universities and research institutes. We believe that communicating the disadvantages of the Forced Swim Test, and providing education on alternative research techniques, will assist in the transition away from the use of the test.⁷⁸

This response shows that, maybe due to practical difficulties, the national regulatory body refrained from advocating a legislative solution against particular animal models, leaving the gatekeeping role to more local actors like ethics committees in universities or research institutes. It is also noticeable that no clear assessment criteria were provided by the regulators.

Conclusion

Since its development, the validity of the FST as an animal model of depression has been challenged, but the test has been resilient and even gradually gained popularity. The past decade of controversy about the FST's legitimacy has had limited effect on its use in academic publications, whereas outside academia the FST has been more heavily contested and even abandoned by many important users. I argue that these distinct outcomes can be explained by the different ways academia and public actors used validity arguments in the legitimacy controversy and the interaction of these arguments with other values.

As shown in the section above titled 'The validity of the FST', academic researchers assessed the FST using Willner's validity framework, consisting of face, construct, and predictive validity. Although most actors applied the same validity framework, different types of validity were prioritized depending on the purpose of the animal model. In research aiming to investigate the pathological mechanism of depression, face and construct validity of

⁷⁷ Transcript segment from video recording, op. cit. (76).

⁷⁸ 'Petition of Tara Jackson', op. cit. (74), p. 4.

the animal model were essential, whereas in research or application for pharmaceutical purposes, predictive validity was seen as of paramount importance. An animal model could have high validity and thus be legitimate to use for one purpose, but lack validity and be illegitimate for a different purpose. In the FST case, it has high predictive validity and thus is widely acknowledged to be a useful tool to screen antidepressants, but its lack of face and construct validity has made the FST a questionable animal model of depression. Since the legitimacy of an animal test depended heavily on its validity for the specific purpose it was purported to serve, changing the target use could influence judgements about its legitimacy. Scientists critical of the FST have mainly focused on its lack of validity for a particular use (i.e. as an animal model of depression), and some of them have gone a step further to propose a new use (i.e. to study stress or coping). When only considering epistemic standards such as validity, the FST could be legitimate for various purposes even though they might be different or in conflict with its original use. While the academic discussion has been more nuanced regarding FST's validity for specific purposes, the public debates driven by animal rights organizations have often taken the academic controversy as evidence of 'invalidity'.

As shown in the section titled 'The validity and legitimacy of the FST', public actors have discussed more explicitly non-epistemic values to explain the FST's current popularity and to advocate its discontinuation. Some pragmatic and social factors could strengthen the FST's legitimacy, resulting in partially cancelling the FST's questioned validity for some purposes. For instance, the FST is known to have many practical advantages, including being inexpensive, easy to set up and use, and quick to produce results. It has also been a well-established animal model that forms part of the tradition in the field. In contrast, ethical concerns weighed in against the FST, with its considerable stress on animals. Animal rights organizations consider the FST's legitimacy as a result of balancing harm and benefit. Facing ethical issues, the FST's legitimacy thus not only requires high validity, but also validity for purposes related to a good enough cause, such as treating depression, which may potentially outweigh the harm the FST procedure may inflict on the animals.

Validity, as an important epistemic value, intertwines with other values, all of which affect the FST's legitimacy. The different presuppositions that actors hold about epistemic, ethical and pragmatic values and their assumptions about the FST's purpose shape how they use validity arguments and determine its legitimacy. Academics like Paul Willner have developed validity frameworks, while critics such as Marc L. Molendijk and Edo Ronald de Kloet have challenged the FST's construct validity and face validity. Meanwhile, activists like Emily Trunnell and organizations like PETA have leveraged these critiques to push for bans, persuading major pharmaceutical companies to abandon the test. Regulatory bodies have responded variably, as seen in New Zealand, where animal rights groups sought a legislative ban, but expert testimony from Craig Johnson led to an emphasis on ethics education instead.

Despite these debates, the question of which validity criteria should take precedence remains underexplored. Should predictive validity alone justify continued use in pharmaceutical testing, or should face and construct validity carry more weight in determining whether the FST is a meaningful model of depression? Furthermore, how should ethical considerations be factored into legitimacy evaluation along with validity assessments? These implicit assumptions often go unspoken, leading to fragmented debates. A clearer legitimacy assessment requires making these presumptions explicit and specifying which forms of validity should be prioritized for particular research objectives. The ongoing controversy surrounding the FST may then gain greater clarity and consistency.

Acknowledgements. I thank the participants of the Validation and Regulation in the Sciences of Health workshop for helpful discussion and feedback on earlier versions of this paper, and I am especially grateful to Lara Keuck and Angela Creager for organizing the workshop and making this stimulating collection possible. I also thank the anonymous reviewers and the editor for their careful and constructive feedback, and one anonymous reviewer in particular for drawing my attention to Harlow's separation studies and their relevance to the early framing of 'despair' in the FST literature. Finally, I am grateful to my supervisors, Henk de Regt, Willem Halffman and Luca Consoli, for their valuable comments and continued support throughout the development of this work.

Competing interests. The author declares none.

Use of artificial intelligence tools. During the preparation of this manuscript, a generative artificial-intelligence tool was used in a limited and supportive manner. Specifically, ChatGPT (OpenAI, large language model accessed via <https://chat.openai.com>) was used intermittently between 2025 and 2026 to assist with language editing. It was not used to generate original ideas, arguments, interpretations, historical claims or references, or to analyse data or source materials. All substantive content and citations were produced by the author, who takes full responsibility for the manuscript, and all AI-assisted text was reviewed and verified prior to inclusion.

Cite this article: Yingying Han, 'Validity arguments on the legitimacy and ethics of the forced swim test in rodents,' *BJHS Themes* (2026), pp. 1–18. <https://doi.org/10.1017/bjt.2026.10035>