Neuroscience and Biobehavioral Reviews 143 (2022) 104928

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

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

Reliability of common mouse behavioural tests of anxiety: A systematic review and meta-analysis on the effects of anxiolytics

Marianna Rosso *, Robin Wirz, Ariane Vera Loretan, Nicole Alessandra Sutter, Charlène Tatiana Pereira da Cunha, Ivana Jaric, Hanno Würbel, Bernhard Voelkl

Division of Animal Welfare, University of Bern, Längasstrasse 120, 3012 Bern, Switzerland

ARTICLE INFO

Keywords:
Mouse
Behavioural tests
Pre-clinical
Anxiety
Anxiolytics
Elevated plus maze
Elevated zero maze
Four-plate test
Holeboard test
Light-dark box
Novelty suppressed feeding
Open field test
Social interaction test
Staircase test
Benzodiazepines
TCAs
SSRIs
SNRIs
Sensitivity
Reproducibility

ABSTRACT

The validity of widely used rodent behavioural tests of anxiety has been questioned, as they often fail to produce consistent results across independent replicate studies. In this study, we assessed the sensitivity of common behavioural tests of anxiety in mice to detect anxiolytic effects of drugs prescribed to treat anxiety in humans. We conducted a pre-registered systematic review of 814 studies reporting effects of 25 anxiolytic compounds using common behavioural tests for anxiety. Meta-analyses of effect sizes of treatments showed that only two out of 17 commonly used test measures reliably detected effects of anxiolytic compounds. We report considerable between-study variation in size and even direction of effects of most anxiolytics on most outcome variables. Our findings indicate a general lack of sensitivity of those behavioural tests and cast serious doubt on both construct and predictive validity of most of these tests. In view of scientifically valid and ethically responsible research, we call for a revision of behavioural tests of anxiety in mice and the development of more predictive tests.

1. Introduction

Animal experiments are a key component of basic and preclinical research, where the mechanisms of diseases are studied and new compounds for their treatment are examined for safety and efficacy before being tested in humans (FDA.gov). However, the use of animals for research can only be justified when the results obtained are informative (Garner, 2014; Würbel, 2017; Henderson et al., 2013), replicable* (Collins and Tabak, 2014; Begley and Ioannidis, 2015; Roth and Cox, 2015), and translatable* (Couzin-Frankel, 2013; Perrin, 2014). Furthermore, public concern for animal welfare urges scientists to comply with the 3Rs principle (Russell and Burch, 1959), that is to refine, reduce, or replace the use of animals whenever possible (Genzel et al., 2020; Directive, /63/EU, 2010). To achieve these goals and ensure responsible scientific practice, the validity* of animal models in use is pivotal (Würbel, 2017; van der Staay et al., 2009; van der Worp et al., 2010; Willner, 1984). A growing body of evidence indicates the lack of validity of animal models as a potential cause for translational failure (Contopoulos-Ioannidis et al., 2003; Geerts, 2009; Howells et al., 2014; van der Worp et al., 2010). Translational failure can slow down medical advancement in the treatment of human disorders (Kola and Landis, 2004; Hay et al., 2014; Leenaars et al., 2019), put patients in clinical trials at risk (Henderson et al., 2013), waste research resources (Olesen et al., 2012), and harm animals for inconclusive research (Russell and Burch, 1959; Würbel, 2017; Bailoo et al., 2014).

Anxiety disorders are amongst the most common mental health conditions, requiring still new and better treatments (Kumar et al., 2013; Harro, 2018; Vos et al., 2016; Kessler et al., 2005; Ravindran and Stein,

* Corresponding author.
E-mail address: marianna.rosso@vetsuisse.unibe.ch (M. Rosso).

https://doi.org/10.1016/j.neubiorev.2022.104928
Received 11 November 2021; Received in revised form 24 October 2022; Accepted 24 October 2022
Available online 29 October 2022

2010). To study anxiety and to test the efficacy of anxiolytic compounds behavioural tests in mice and other animals are commonly used (Crawley, 2007; Kumar et al., 2013; Harro, 2018; Belzung and Griebel, 2001). Behavioural tests for anxiety are either based on conditioned fear, acute fear responses, or on exploiting an approach-avoidance conflict (Haller and Alicki, 2012). In the latter case, the conflict an animal may experience between exploring a new, and avoiding a potentially threatening, environment is assumed to elicit measurable changes in the behaviour of the animal (Ennaceur, 2014; Crawley, 2007; Hånell and Marklund, 2014). Amongst the various behavioural tests for rodents, the open-field test is arguably the most popular one (Harro, 2018). This test, although with several modifications (Walsh and Cummins, 1976; Belzung, 1999), generally consists of a brightly illuminated arena, enclosed by walls. During the test, an animal is placed inside the arena and behavioural outcomes are recorded. The test was originally established to assess emotionality in rats, using urination and defecation as measures of timidity (Hall, 1934; Walsh and Cummins, 1976). The use of the open-field test was then extended to assess a wider range of behavioural features and psychiatric conditions (Crawley, 2007) and adopted for other species. Similar to rats, early studies which employed the open-field test in mice measured defecation and freezing to assess genetic differences in behaviour (DeFries et al., 1974, 1966). Additionally, the distance travelled in the open-field test has been introduced and - since then - widely used as a measure of locomotor activity to assess, for instance, the effect of sedative or stimulant drugs (Prut and Belzung, 2003), as well as anxiety expressed by the animals (Crawley, 2007; DeFries et al., 1974). Overall, anxiety is thought to suppress animals' drive to explore a novel environment, while vice-versa, less anxious animals are predicted to explore comparatively more. In the open-field, exploration has been previously quantified by measures such as total locomotion or number of vertical rearing (Ohl et al., 2008; Harro, 2018; Crawley, 2007). Further, thigmotaxis in the open-field, namely the tendency to stay in proximity of the walls while avoiding the centre of the arena, is often recorded and interpreted as a proxy for anxiety (Bourin et al., 2007; Belzung, 1999; Crawley, 2007).

Similar to the open-field test, the elevated plus maze test (Montgomery, 1958) and the light-dark box test (Crawley and Goodwin, 1980) are based on the conflict between the exploration of a new environment and the natural aversion of rodents to bright and open spaces. Overall, the rationale behind these tests as measures of anxiety rests on the assumption that a state of anxiety should modulate the animals' behaviour by reducing exploration, therefore reducing the exposure to (potential) threats (Kumar et al., 2013; Crawley, 2007; Ohl et al., 2008). Accordingly, the efficacy of anxiolytic compounds is assessed based on whether and to what extent they attenuate the reduction of exploratory behaviour by the test situation. Other popular tests, such as the hole-board test (File and Wardill, 1975), the elevated zero maze (Shepherd et al., 1994), the social interaction test (File and Hyde, 1979), the novelty suppressed feeding test (Bodnoff et al., 1988), and the four-plate test (Aron et al., 1971), are based on the same conceptual rationale.

Over the years, behavioural tests for anxiety have been considered validated, because of reported behavioural changes elicited by benzodiazepines, and specifically diazepam (Ennaceur and Chazot, 2016; Bespalov and Steckler, 2021; Cryan and Sweeney, 2011). However, anxiolytic agents such as benzodiazepines also possess anti-depressant and sedative effects, which implies that the observed behavioural effects may not necessarily be due to a change in anxiety, but could be a result of the sedative properties of the drug (Prut and Belzung, 2003).

Despite their popularity, several experimental studies, as well as literature reviews, have highlighted inconsistent results in the behavioural outcomes elicited by new classes of anxiolytics, therefore questioning the suitability of these outcomes as indicators for anxiety (Ennaceur and Chazot, 2016; Ennaceur, 2014; Prut and Belzung, 2003; Hascoët and Bourin, 1998; Rodgers et al., 1997; Haller and Alicki, 2012). Benzodiazepines, although popular in the past to treat anxiety, have now been replaced by better pharmacological compounds with fewer side effects and lower withdrawal-related risks (Bystrisky et al., 2013; Costa et al., 2014; Moniruzzaman et al., 2018). Selective Serotonin Reuptake Inhibitors (SSRIs) or Serotonin–Norepinephrine Reuptake Inhibitors (SNRIs), which are now used as a first-line pharmacological treatment for human anxiety disorders, have failed to give reliable results in rodent behavioural tests of anxiety (Rodgers et al., 1997; Prut and Belzung, 2003; Ennaceur, 2014; Ennaceur and Chazot, 2016; Borsini et al., 2002).

Here, we aimed to assess the validity of common behavioural tests of anxiety in mice by evaluating their responsiveness to anxiolytic compounds prescribed to humans, a process known as 'reverse translation' (Hart, 2015; Shakhnovich, 2018). To this end, we performed a pre-registered systematic review of research papers that had used these tests on laboratory mice, for a broad range of anxiolytic compounds. We investigated the overall effect size for a range of test measures of common behavioural tests as well as the variation of the reported outcomes across the published literature. Additionally, we evaluated sample heterogeneity and estimated the quality of reporting through a risk of bias assessment.

***Glossary of key terms**.

1. Replicability: Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

- o Relevant literature: (Baker, 2015; Begley and Ioannidis, 2015; Freedman et al., 2015; Roth and Cox, 2015; Baker, 2016; Miyakawa, 2020; Smith and Lilley, 2019)

1. **Translatability**: the extent to which results obtained in an animal model can be replicated in the system which is being modelled.
   - o Relevant literature: (Hackam and Redelmeier, 2006; Geerts, 2009; Mak et al., 2014; Kola and Landis, 2004; Howells et al., 2014; O'Collins et al., 2006)
2. **Validity**: to be fit for use in research, and therefore be considered to be a valid animal model, a test or animal model should meet several criteria of validity, including:
   - i. **Construct validity**: the extent to which the test can measure what it is supposed to measure
   - ii. **Predictive validity**: the extent to which a test can predict a certain outcome in the system that is being modelled.
   - o Relevant literature: (Belzung and Griebel (2001); Würbel (2017); van der Staay et al. (2009); Garner et al. (2017); Garner (2014); Steimer (2011)

2. Results

2.1. Study selection

Our search retrieved 744 papers from PubMed and 2533 papers from Embase of which 1764 were excluded in the first steps of the review (Fig. 1). In particular, 533 were excluded as paper duplicates, and 1231 were excluded based on abstract and/or method section screening. The full texts of 1513 papers were screened and 814 of those papers were included in the data extraction process according to the pre-specified criteria. As the search strategy identified key words in all fields of the text, several papers not relevant to us were identified; 62 papers were excluded because the text was unavailable publicly, 331 papers were excluded because the sample size was unclear or not reported, 64 papers were excluded due to incompatible outcomes reported (for instance, the correct test and drugs were used, but none of our outcomes of interest were reported), 59 papers were excluded because compounds other than the ones of interest were used, or compounds were used in combination with other compounds, 48 papers were excluded because of issues in the reporting of the outcomes, 40 papers were excluded because they had formats other than research papers, 33 papers were excluded because

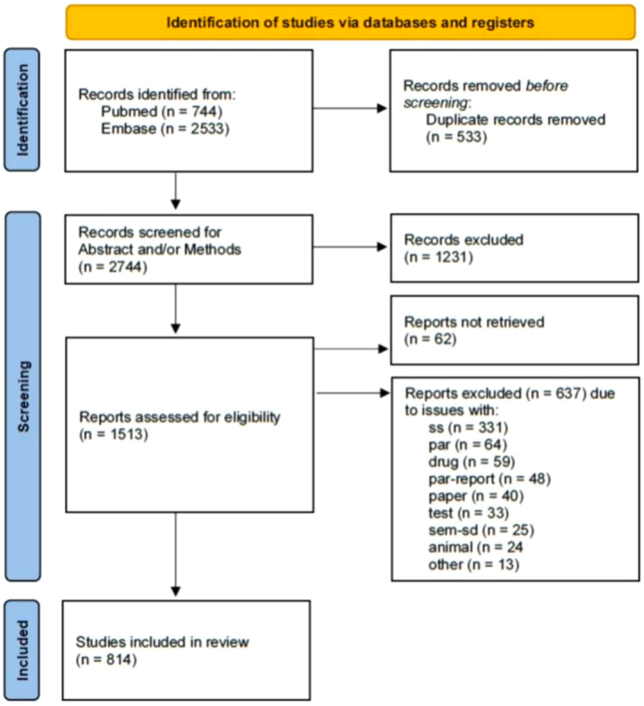

Fig. 1. Flowchart of the screened papers and reasons for exclusion. ss: unclear or absent sample size; par: incompatible outcomes reported; drug: incompatible compounds used; par-report: issues with the reporting of the outcomes; paper: wrong format of paper; test: incompatible behavioural test used; sem-sd: unclear or absent measure of variance; animal: wrong animals used; other.

the behavioural tests used were different from the ones of interest, 25 papers were excluded due to ambiguity regarding the measure of variance of the reported outcomes, 24 papers were excluded because they used animals other than mice, or because of ambiguity in the species of animal used, and 13 papers were excluded for other reasons (i.e. missing controls, treatment administered to mothers, etc.).

2.2. Study characteristics

All the eligible studies used mice, which were tested in behavioural tests after administration of anxiolytic compounds. The Supplementary table illustrates the details of data distribution in the different test measures of interest in combination with each compound. Due to reporting of multiple outcomes per paper, a total of 2476 outcomes were distributed across 17 different test measures, in combination with 25 different anxiolytic compounds. The test measures from the elevated plus maze and the open field made up the great majority of outcomes (74 %, Table 4), followed by the light-dark box test and the holeboard test contributing a total of 13 % and 5 % of the outcomes, respectively. A minor contribution was attributed by the staircase test (the staircase test, n = 56, "rrs" n = 27, "stps" n = 29), the four-plate test (n = 42), the novelty suppressed feeding test (n = 37), the social interaction test (n = 26), and the elevated zero maze (n = 14). The great majority of these measures were recorded when used in combination with benzodiazepines (72 %), with diazepam being the most frequently used compound (65 %), mainly with a dosage of either 1 mg/kg or 2 mg/kg (56 % and 31 % of the cases respectively). SSRIs was the second most common compound class (20 %), with fluoxetine (12 %), either 10 mg/kg (in 34 % of the cases) or 20 mg/kg (in 31 % of the cases), being its most frequently used representative. Further information regarding the dosages used in the study are reported in the supplementary material.

Table 1
Results of the risk of bias assessment. Values in the table indicate percentages of papers, which scored either as high, medium, or low risk of bias in each item (row).

Question	High	Medium	Low
was an automatic randomisation method used to allocate animals to groups?	97.22	2.78	0
were animals randomly allocated to treatment/control group?	65.56	34.44	0
was the test order randomised or counterbalanced?	92.78	6.11	1.11
was the sample size declared to be appropriately calculated?	98.89	1.11	0
where animals randomly housed?	95.56	4.44	0
compliance with animal welfare regulations declared?	19.44	48.89	31.67
were the investigators blinded during the experiment?	95.56	3.89	0.56
is the statistical analysis described?	2.22	54.44	43.33
Is the housing temperature reported?	25	0	75
Is the sex of the animals reported?	10	0	90
Is the strain of the animals reported?	5	0	95
conflict of interest declaration	52.78	0	47.22
publication in a peer-reviewed journal?	0	0	100
were the outcome assessors blinded during the experiment?	65.56	0	34.44

2.3. Risk of bias

A sub-sample of 180 papers was analysed in detail to assess risks of bias across 17 different items (Table 1). All the scored papers were published in peer-reviewed journals, and most of them reported mouse strain (95 %), sex (90 %) and housing temperature (75 %). 31 % of the papers reported details regarding compliance with animal welfare regulations, 43 % of the papers reported details on the statistical analysis, and 34 % of the papers reported details on the blinding procedures. For the following five items, we scored a high risk of bias: automatic allocation to treatment group (97 %), randomised order of testing (92 %), a-priori sample size calculation (98 %), random housing (95 %), and blinding of investigators (95 %). Further details are reported in Table 1.

2.4. Synthesis of results

Estimated effect sizes varied greatly across the majority of the test measures and compounds (Fig. 2). The overall estimated effect size allows determining whether there is evidence of an anxiolytic effect on the behavioural measures elicited by a range of anxiolytic compounds. Ten out of the 17 test measures yielded a positive overall effect size significantly different from zero (EPM-*eca*, EPM-*eoa*, EPM-*toa*, FPT-*cross*, LDB-*light*, LDB-*trans*, NSF-*lat*, OF-*cent*, SI-*time*, STC-*rrs*), while overall effects of the remaining seven did not significantly deviate from zero.

For each meta-analysis, the factor 'compound' was tested for significance to assess whether any of the anxiolytic compounds affected behavioural outcomes. For this, the null hypothesis to be tested assumes the estimated effect sizes for all compounds to be zero (Viechtbauer, 2010). After family-wise correction for multiple testing for the 17 meta-analyses performed, five measures showed no significant effect, namely EZM-*toc*, LDB-*dark*, NSF-*lat*, OF-*dist*, and SI-*time* (Table 2).

For each test measure, we calculated total and partial I^2 as a measure of heterogeneity. For 15 out of 17 measures, total I^2 was above 85 %. The partial I^2 attributed to 'strain' contributed little to the total I^2, except for SI-*time*, where it accounted for 48 % of the total heterogeneity. Similarly, administration mode (acute or chronic) had in all cases either no or only small effect (<10 %). Partial I^2 attributable to within-study heterogeneity varied greatly across measures: in 10 cases being < 10 %, while being more pronounced in others (e.g. 64 % for FPT-*cross*). Between-study heterogeneity explained the greater part of the total heterogeneity for 14 of the 17 measures (Table 2).

Given the 25 compounds and 17 test measures, there are a total of

Table 2
Abbreviations for tests and measures as listed in the methods section. Factor compound gives the significance of moderator effect (treatment × compounds interaction), total and partial I^2 estimates per test measure.

Test	Measure	Factor 'compound'	I^2 Total	I^2 between studies	I^2 within study	I^2 Strain	I^2 Chronicity
EPM	eca	ns	90.9	80.0	3.3	0.5	7.1
	eoa	*	87.5	57.2	8.4	20.6	1.3
	toa	*	94.3	73.4	4.5	16.4	0.0
EZM	toc	ns	85.7	0.0	0.0	82.8	2.9
FPT	cross	*	85.5	21.5	63.9	0.0	0.0
HBT	hd	*	97.7	97.7	0.0	0.0	0.0
LDB	dark	ns	99.2	99.1	0.1	0.0	0.0
	light	*	96.3	92.5	0.7	3.1	0.0
	trans	*	72.3	67.4	0.0	4.9	5.1
NSF	lat	ns	91.8	54.9	36.9	0.0	0.0
OF	cent	*	90.6	73.2	10.2	0.0	7.1
	dist	*	82.9	54.9	19.7	6.8	1.5
	rear	*	93.5	88.2	1.9	0.6	2.7
	sqrs	*	95.1	84.5	7.8	0.6	2.1
SI	time	*	94.6	0.0	45.9	48.7	0.0
STC	rrs	*	86.1	59.9	26.2	0.0	0.0
	stps	*	97.1	78.5	0.0	17.9	4.1

Source:Sources of heterogeneity in the study sample.

Table 3
Number of studies and percentage of positive studies, per combination of test measure and anxiolytic compounds. Cells in grey indicate a percentage of positive studies < 75 %. Coloured cells highlight a percentage of positive studies > 75 %. Colour gradient indicates an increasing number of studies. Combinations with only one study were excluded from the table.

		Tests	EPM			EZM	FPT	HBT	LDB			NSF	OF				SI	STC	
		Measures	eca	eoa	toa	toc	cross	hd	dark	light	trans	lat	cent	dist	rear	sqrs	time	rrs	stps
Benzodiazepine	Alprazolam	*n*		4	8		4			4	2				2	3			
		% sign.		75%	75%		50%			100%	50%				100%	33%			
	Chlordiazepoxide	*n*	21	17	24	4		2		8	7	2				triazolam		2	2
		% sign.	5%	59%	71%	50%		0%		63%	29%	100%						50%	0%
	Clorazepate	*n*									2			clonazepam		2			
		% sign.									100%					0%			
	Diazepam	*n*	138	221	413	4	34	120	27	142	84	3	37	28	121	207	14	19	21
		% sign.	38%	82%	84%	100%	74%	35%	67%	80%	62%	67%	59%	18%	21%	25%	71%	79%	48%
	Lorazepam	*n*			2					3	2				2	2			
		% sign.			50%					100%	100%				50%	100%			
Other	Hydroxyzine	*n*	2	2	3					5			trazodone		2	2			
		% sign.	100%	100%	100%					60%					100%	100%			
SNRI	Duloxetine	*n*	2	2	4				2	4						2			
		% sign.	0%	100%	0%				50%	50%						50%			
	Venlafaxine	*n*			3								2	5	3	5			
		% sign.			33%								0%	0%	33%	20%			
SSRI	Buspirone	*n*	22	25	32					3				3	3	6			
		% sign.	32%	20%	34%					100%				0%	33%	17%			
	Citalopram	*n*			3					2			6	9		3			
		% sign.			0%					100%			0%	44%		33%			
	Escitalopram	*n*	2		5			4				3	2	2		6			
		% sign.	0%		40%			0%				0%	0%	0%		0%			
	Fluoxetine	*n*	13	14	35			5		10	4	21	27	56	43	69	9		
		% sign.	8%	21%	23%			60%		50%	0%	19%	30%	13%	26%	17%	44%		
	Fluvoxamine	*n*											2		3	5			
		% sign.											0%		67%	0%			
	Paroxetine	*n*			3					2			3			6			
		% sign.			0%					100%			33%			0%			
	Sertraline	*n*			3										3	4		2	2
		% sign.			67%										33%	25%		100%	0%
TCA	Amitriptyline	*n*												2		3			
		% sign.												50%		0%			
	Imipramine	*n*	4	5	9			5		3	2	2	3	17	18	32	2		
		% sign.	0%	60%	56%			80%		100%	0%	100%	33%	18%	28%	25%	50%		
	Nortriptyline	*n*			2								maprotiline		2	2			
		% sign.			50%										100%	100%			

425 compound-by-measure combinations. We found reported study outcomes for 182 of those compound-by-measure combinations (details summarised in the Supplementary Table). The number of outcomes per combination varied from 1 to 413, with 118 compound-measure combinations with more than one outcome recorded. Of these, only 32 had a positive and significant effect size (i.e. the lower bound of the 95 % confidence interval being larger than zero), while 86 combinations did not show a positive effect (Fig. 2 and Supplementary Table). Diazepam was the compound that elicited a significant positive effect size in 9 out of 17 test measures. Overall, most of the combinations with a significant effect size were due to benzodiazepines, with 20 positive effects out of 32. LDB-*light* yielded a positive effect size for most of the anxiolytic

Table 4
Behavioural tests for anxiety in mice and relative test measures included in the search.

Test	Test measure	N Outcomes retrieved	Included
Elevated plus maze (EPM)	eca: Number of entries into closed arms.	206	yes
	eoa: Number of entries into open arms.	296	yes
	toa: Time (both in percentage and in time unit) spent in the open arms.	552	yes
Elevated zero maze (EZM)	ecc: Number of entries into the closed compartment.	2	no
	eoc: Number of entries into the open compartment.	5	no
	toc: Time (both in percentage and in time unit) spent in the open compartment.	14	yes
Four-plate test (FPT)	cross: Number of punished crossings.	42	yes
Holeboard test (HBT)	hd: Number of head dips.	137	yes
Light-dark box (LDB)	dark: Time spent in the dark compartment.	35	yes
	light: Time (both in percentage and in time unit) spent in the light compartment.	187	yes
	trans: Number of transitions between the two compartments.	107	yes
Novelty suppressed feeding (NSF)	lat: Latency to eat (sec).	37	yes
Open field test (OF)	cent: Time (both in percentage and in time unit) spent in the center (as defined by the authors).	87	yes
	dist: Distance travelled.	125	yes
	rear: Number of rearings.	207	yes
	sqrs: Number of squared crossed.	362	yes
Social interaction test (SI)	time: Time (sec) spent in social interaction.	26	yes
Staircase test (STC)	rrs: Number of rearings.	27	yes
	stps: Number of steps climed.	29	yes
Vogel conflict test (VC)	dbs: Number of drinking bouts.	7	no
	shck: Number of shocks accepted or received.	9	no

compounds tested, 8 out of 11, and EPM-*toa* yielded a positive effect size for 5 out of 15 anxiolytic compounds. The rest of the test measures detected an effect for at most two anxiolytic compounds, across the range with which they were tested.

The percentage of individual observations that detected a positive significant effect varied greatly across the different combinations of test measures and anxiolytic compounds, ranging from 0 % to 100 % (Table 3). Overall, only 1254 of all 2476 contrasts (i.e. 50 %) showed significant positive treatment effects. As all the compounds included in this analysis have been shown to reduce anxiety in humans, we assessed the sensitivity of behavioural tests outcomes to detect the expected anxiolytic effect of these compounds in mice based on the logic of reverse translation. Thus, we used the proportion of individual studies reporting a significant positive effect as a measure of sensitivity and an estimate of the true positive rate. To conclude that a behavioural test reliably detects an anxiolytic effect, we require that individual studies detect significant effects (positive effect size with a 95 % confidence interval not including zero) in at least three out of four cases (i.e. 75 %). The majority of behavioural measures failed to reliability detect an effect for the majority of the compounds. In 89 out of 118 combinations for which more than one outcome was recorded, less than 75 % of individual studies reported significant positive effects, while only for 29 combinations, the proportion was greater than 75 %. Table 3 suggests that diazepam was the compound that most often elicited a behavioural change detectable in five test measures. Here, we also observe a higher number of studies as compared to other compounds. Out of the 29 'reliable' combinations, benzodiazepines were the dominant compound class, showing reliable results in 14 combinations. LDB-*light* seems to be the most promising candidate to detect an anxiolytic effect, with the majority of individual studies detecting an effect in seven out of 11 anxiolytic compounds across compound classes. Furthermore, EPM-*eoa* and EPM-*toa* reliably detected effects for 3 and 4 anxiolytic compounds, respectively. Similarly, OF-*sqrs*, reliably detected an effect of 3 anxiolytic compounds, but the number of individual studies was far lower than for the EPM. Forest plots (Fig. 3 and Supplementary Material) show how for some measures the estimated effect sizes for individual studies range from highly negative values to highly positive ones, spreading in an almost symmetrical fashion across the null. Clear examples of such pattern can be seen in the forest plots of EPM-*eca*, HBT-*hd*, LDB-*trans*, NSF-*lat*, OF-*dist*, OF-*rear*, OF-*sqrs*, and STC-*stps*.

3. Discussion

With the present study, we aimed at providing a synthesis of the reliability of the most commonly used behavioural tests of anxiety, as they are currently performed. Mice are by far the most commonly used species in biomedical research, as reported, for instance, by the latest statistical report on the animals used in research in the EU (Directive, /63/EU, 2010). Here, it was reported that nearly 200,000 mice were used in 2018 for translational research on human nervous and mental disorders (i.e. 63 % of the total animals used in this field). Given their widespread use and the resulting concern in term of animal welfare, and medical and scientific relevance, the optimisation of the tools used and the tests performed represents an important step in medical and scientific advancement as well as a critical animal welfare concern. Hence, our study focused on behavioural tests for mice. We assessed their sensitivity to a broad range of anxiolytic compounds approved for the treatment of anxiety in humans, using a systematic and unbiased approach. Briefly, we found reported effects to vary greatly across studies and test measures, in addition to overall high heterogeneity and substantial risks of bias based on how studies are reported.

We found that for five of the 17 test measures, none of the anxiolytic compounds had a significant effect, whereas, for the remaining 12 test measures, an effect of at least one anxiolytic compound was detected. Additionally, we investigated the overall estimated effect size for each test measure, irrespective of anxiolytic used, and found null or negative overall effects for seven test measures.

For the majority of the test measures and anxiolytic compounds, we observed great variation in the estimated effect sizes, ranging from highly negative to highly positive values, and resulting in estimated cumulative effect sizes close to zero (e.g. in NSF-*lat*, STC-*stps*, and in HBT-*hd*.). Additionally, we observed that the effect size estimates of individual studies reporting a significant effect of a compound also varied greatly, even for combinations of test measure and compound for which the estimated cumulative effect size was positive. Since all of the compounds included in our study were shown to have anxiolytic effects in humans, we consider the proportion of individual studies reporting an effect as a measure of how reliably these behavioural test measures can detect behavioural changes elicited by anxiolytic compounds.

Analysis of the total and partial heterogeneity showed that across test measures the greater portion of the sample heterogeneity was produced by differences between studies. Such a high level of between-study heterogeneity seems to be common in various fields of animal research (Pires et al., 2016; Antoniuk et al., 2019; Leffa et al., 2019; Voelkl et al., 2018).

There were only two test measures for which between-study heterogeneity was as low as expected due to random variation alone: SI-*time* and EZM-*toc*. These test measures were, however, not sensitive to effects of anxiolytic compounds. Within-study heterogeneity varied greatly

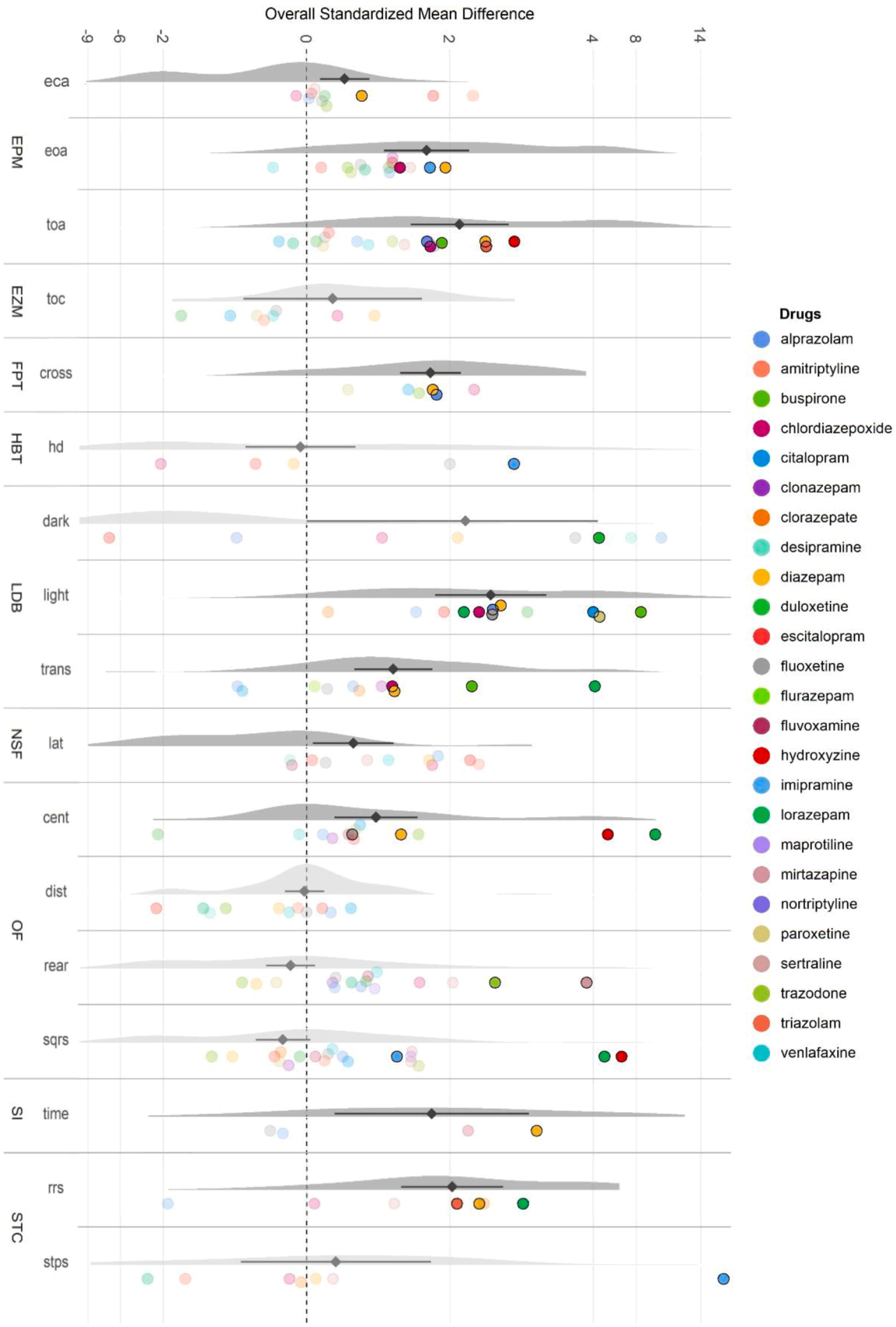

Fig. 2. Distribution plots showing the probability density distribution of the calculated effect size (x-axis) of the individual studies for each test measure. The overall estimated effect size for each test measure, is indicated by a diamond, and the related 95 % confidence interval by a horizontal line. Points indicate the estimated mean effect size for each compound. Colours indicate anxiolytic compounds. Opacity is applied to not significant effect sizes, i.e. the lower bound of the 95 % confidence interval is lower than zero. An interactive version of the figure can be found online at https://mrossovetsuisse.shinyapps.io/Shiny_SR/.

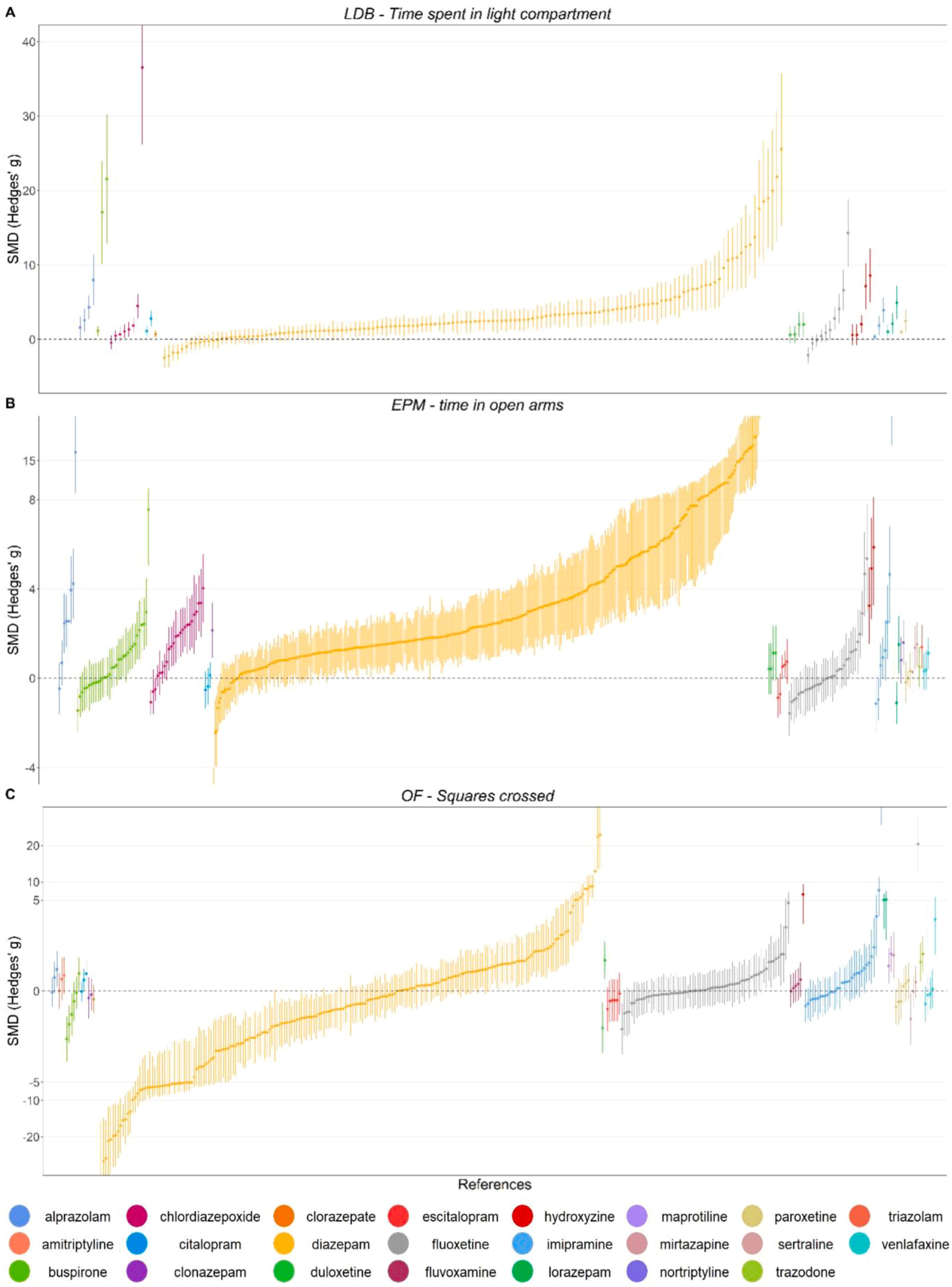

Fig. 3. Forest plots of three selected test measures: A: LDB-*light*, B: EPM-*toa*, C: OF-*sqrs*, sorted for increasing effect size. Different colours indicate different anxiolytic compounds, as indicated in the legend (See Supplementary material for remaining measures).

across measures but was overall lower than other measures of partial heterogeneity, hinting at high levels of standardisation within laboratories (Voelkl et al., 2018).

Even though our results reveal that most of the common test measures of anxiety do not reliably detect behavioural changes elicited by several anxiolytic compounds, we have found two test measures - EPM-*toa* and LDB-*light* - that appear to be sensitive in terms of detecting both a positive effect for most anxiolytic compounds and a positive effect in the majority of individual studies. With 73 % (EPM-*toa*) and 78 % (LDB-*light*), respectively, of individual studies reporting a positive effect, the false-negative rates approach the minimally recommended threshold of 0.2. Though, as those numbers show, even for those tests refinements that would increase their reliability would be highly desirable.

The substantial variation observed among studies using the same test measure and the same anxiolytic compound with comparable dosages is likely to be attributed to environmental, genetic, and procedural differences. Previous analyses of behavioural test outcomes for the effect of mouse strain on both basal levels of performance and performance after the administration of anxiolytic compounds highlighted substantial strain differences and often conflicting results (Ennaceur and Chazot, 2016; Griebel et al., 2000; Bouwknecht and Paylor, 2002; Hagenbuch et al., 2006; Gard et al., 2001). Surprisingly, we found only weak effects of mouse strain on heterogeneity for most test measures. This finding is, however, in line with the results of recent meta-analysis of the effects of benzodiazepines and SSRIs on the behaviour in the marble burying test, where the authors also found no indications for strain effects (Langer et al., 2020). Apart from genetic background, differences in sex, age, housing conditions, and test environment and procedure may contribute to between-study variation. Unfortunately, such information is only sporadically and scantily reported. In our sample, only an average of 5 % of the animals used were females, which renders the analysis of sex influence challenging. Key aspects of test protocols are rarely reported which does not allow for an in-depth analysis of the test procedures.

Whether a compound is administered chronically or acutely may have an effect of the resulting outcome. In fact, for compounds such as SSRIs chronic treatment is required to achieve the desired effect (Benmansour et al., 1999). Most of the studies included in the present review (68 %) reported acute administration, while only 16 % of the studies reported chronic administration, though similar patterns for the distribution of effect sizes were found either case and the meta-analyses suggest that administration modality accounts only for a small percentage of the overall heterogeneity. We invite the readers to visit our online application, available at https://mrossovetsuisse.shinyapps.io/Shiny_SR/, which allows exploring our dataset by sex, strain, stress treatment, and drug dosage.

Taken together, our results show that most commonly used behavioural test measures of anxiety are unreliable in detecting behavioural changes elicited by anxiolytic compounds other than benzodiazepines and in particular diazepam. This corroborates the previously voiced suspicion that most popular behavioural tests of anxiety (with the exception of EPM-*toa* and LDB-*light*) are in fact "benzodiazepines tests" (Bespalov and Steckler, 2021; Ennaceur, 2014). The behavioural effects elicited by benzodiazepines in these tests have been proposed to reflect disruption of normal behaviour, possibly resulting in altered impulse control rather than attenuated anxiety (Bespalov and Steckler, 2021; Thiébot et al., 1985).

The behavioural tests included in our study heavily rely on changes in exploration patterns to determine anxiety levels and such test procedures may not be able to disentangle behavioural changes in exploration and anxiety (Bourin et al., 2007; Hascoët and Bourin, 1998; Andreatini and Bacellar, 2000). A clear example of this problem is the open field test. Here, we identify an issue with the continuation of such tests, as long-held standard that may not be appropriate, due to the researcher's degrees of freedom in the interpretation of the test's meaning (Wicherts et al., 2016; Pound and Bracken, 2014). On the other hand, behavioural tests not included in the present review which do not rely on changes in exploratory behaviours such as the marble burying test, may represent promising alternatives for the assessment of anxiety in mice (Haller and Alicki, 2012; Langer et al., 2020; Broekkamp et al., 1986).

On a different note, our findings question the standard classification of effect sizes in animal behavioural research. Cohen introduced what are, up to date, considered the conventional thresholds for small, medium, or large effect sizes (namely, a Cohen's d of 0.2, 0.5, and 0.8 respectively (Cohen, 1977)). The author warned for caution (p. 25) in using these thresholds for power analysis outside the scope of the field for which they were initially thought for (psychology or sociology). Study populations of laboratory animals are normally characterised by high degrees of both genetic and environmental standardisation (Wahlsten et al., 2003; Wahlsten, 2001; Würbel, 2000). Therefore, populations of animal studies are usually much more homogenous, producing much lower levels of random variation, when compared to study much more heterogenous populations of clinical studies, resulting in higher within-study variation, but lower between-study variation and therefore smaller heterogeneity (Voelkl et al., 2020, 2018). This difference has important implications for the interpretation of standardised effect sizes like Cohen's d or Hedges' g. Due to the higher level of standardisation in animal studies, as compared to clinical studies, and the resulting low within-group variation, a given mean difference between a control and a treatment group will result in a much higher standardised effect size. For example, for EPM-*toa,* (Santana et al., 2014) reported 123.8 s spent in the open arms for the control group and 207.3 s for the group receiving diazepam. Given the corresponding standard errors of 0.4 and 0.7 for the control and the treatment group, respectively, this amounts to a standardised effect size of 40.6, which is on an entirely different scale of magnitude than a Cohen's d of 0.8, the reference for "large" effects. While this is one of the more extreme examples, we note that EPM-*toa* had an average effect size across drugs of 2.13, with 77 % of the total studies reporting an effect size larger than the standard large effect of 0.8. Correct estimation of expected effect sizes is essential for proper power analyses and sample size calculations, with important implications for animal welfare. Considering the large achieved effect sizes, the power analyses based on the "standard Cohen's values" are likely to lead to unnecessarily large required sample sizes. Because of this, we call for a cautious interpretation and more contextualised use of effect size classification, according to each field of research.

Risk of bias assessment showed overall high-risk scores for most of the items. Although the common checklists and tools for risk of bias analyses assess reporting quality rather than study quality, high risks of bias can have serious implications for the reproducibility and replicability of study findings. Albeit efforts have been made to develop more stringent guidelines for both designing and reporting of animal studies (Du Percie Sert et al., 2020; Smith et al., 2018), we observed an overall low quality of reporting, which likely reflects poor study design and conduct. For instance, important aspects of the housing conditions (e.g. temperature), randomisation and blinding procedures, testing conditions (e.g. apparatus size, light intensity, and time of testing), and sample size calculations were reported only sporadically, and in 10 % of the cases researchers even failed to report the sex or the strain of the animals.

A risk of bias analysis suggested overall high risk of bias for a large majority of the publications, reflecting a very succinct and partly patchy reporting practise in this field. As studies were published in a variety of journals, we explored whether results differed depending on the journal impact factor where the study was published, though we found no overall differences (supplementary material).

Finally, we want to issue a caveat. In systematic reviews, researchers commit to a pre-defined and pre-registered search strategy. This commitment reduces selection bias, because researchers cannot 'adjust' their sample retrospectively. However, strict adherence to search terms does also mean that relevant studies can be missed only because of different wording. This is unavoidable in systematic reviews, though it is

not a problem if one keeps in mind that the purpose of a systematic review is not to produce an exhaustive list of all studies but a representative, unbiased sample of studies. In our case, the final sample of 814 included studies is, for sure, smaller than the total number of studies that used behavioural tests of anxiety. Our sample does also not contain all test and test measures ever used. For example, we missed several studies using the marble burying test because we searched for "defensive marble burying test". Our study contains, however, the most commonly used tests and test measures and a representative, unbiased sample of those tests.

In conclusion, our findings indicate that most common behavioural measures of anxiety in mice have low to no sensitivity to anxiolytic compounds commonly used for the treatment of anxiety in humans. This is especially true for compounds other than diazepam. These findings further suggest that most of these test measures also have poor predictive validity for the discovery of new compounds to treat anxiety disorders in humans and point at a high false-negative rate for individual studies. 382 studies in our sample (47 %) relied exclusively on measures for anxiety for which our meta-analysis delivered no indication that those tests are fit for purpose. This means that in our sample alone 11880 animals were used in tests with no predictive validity. Additionally, we observed considerable idiosyncrasy in the results of individual studies, with effect sizes ranging from highly negative to highly positive for most outcomes and most compounds suggesting that most behavioural tests, as they are currently performed and interpreted, produce unreliable and often even contradicting results. These findings are corroborated by previous evidence for poor replicability of behavioural tests for anxiety (Bespalov and Steckler, 2021; Ennaceur and Chazot, 2016). Poorly validated animal tests that produce idiosyncratic results fail to generate new knowledge and, consequently, may lose their ethical justification. Following the 3Rs principle, efforts must be made to improve the validity of animal tests of anxiety.

4. Methods

4.1. Pre-registration

Prior to data extraction, in November 2019, this study was pre-registered at SYRCLE (see supplementary information for the pre-registration protocol).

4.2. Search strategy

The search strategy consisted of i) a list of anxiolytic compounds, ii) the keyword “mice”, and iii) a list of behavioural tests for anxiety. To define the list of anxiolytic compounds, we used a combination of the following databases to list compounds that are commonly used to treat anxiety disorders in humans: DrugBank (drugbank.ca); FDA Drug Approval Databases (FDA.gov); Anxiety and Depression American Association (adaa.org). We selected the following compounds: alprazolam, amitriptyline, buspirone, chlordiazepoxide, citalopram, clomipramine, clonazepam, clorazepate, desipramine, diazepam, doxepin, duloxetine, escitalopram, fluoxetine, flurazepam, fluvoxamine, hydroxyzine, imipramine, lorazepam, maprotiline, mirtazapine, nortriptyline, oxazepam, paroxetine, protriptyline, sertraline, temazepam, trazodone, triazolam, trimipramine, venlafaxine. A literature search allowed us to identify behavioural tests commonly used to assess anxiety in mice, as well as the most commonly reported behavioural outcome measures (Table 4). Each test that yielded more than 10 results, when searched on PubMed (on date July 15th 2019) in combination with the aforementioned list of compounds, and the keyword “mice”, was included in the search (Supplement 1). A minimal threshold of 10 outcomes was chosen to exclude measures with insufficient evidence to be analyzed. A detailed list of the behavioural tests searched and included can be found in the Supplement 1. The search was performed on PubMed (ncbi.nlm.nih.gov/pubmed) and Embase (embase.com), on August 21st, 2019.

4.3. Study selection

After reference retrieval, we excluded paper duplicates using the reference manager software Citavi 6.4 (Swiss Academic Software GmbH, Wädenswil, CH). The main reviewer (MR) scanned the titles, abstracts and/or methods of these papers, and excluded all those, which did not use the behavioural tests of interest (Table 4), mice, or the selected anxiolytic compounds. Additionally, we excluded papers that were not original research papers and papers that were not written in English. After the first scan, each paper was fully screened independently by two reviewers (main reviewer: MR, second reviewers: RW, AL, NS) who also independently extracted the data. Discrepancies between the two reviewers were resolved through discussion, or with the aid of an external reviewer (BV).

4.4. Study characteristics

Studies were included or excluded according to the pre-specified inclusion/exclusion criteria (Supplement 1). For each paper, two reviewers independently extracted information about the animals (i. strain, ii. sex, iii. age, iv. transgenic ID; v. stress or defeat treatment), about the treatment (vi. compound, vii. dosage, viii. route of administration, ix. time of administration before testing), and about testing (x. open field size, xi. test duration). For each test, we selected test measures suggested by the authors as measures of anxiety (Table 4). For each test measure, we extracted mean values, sample size, and either standard deviation or standard error of the mean, for both treatment and control group. We accepted any control group as declared by the authors (e.g. administrating water, saline solution, etc.). Information from graphical data was extracted using the online software Automeris (https://apps.automeris.io/wpd/).

4.5. Data analysis

The statistical analysis was performed in R (1.4.1103) (R Core Team, 2020) with the package metafor 2.4–0 (Viechtbauer, 2010). For each study, we computed the standardised mean difference Hedges’ g between the control and the treatment group as the chosen indication of effect size (metafor::escalc). We included any test measure that yielded at least 10 results. Consequently, four measures (EZM-*eoc*, EZM-*ecc*, VT-*shcks*, VT-*dbs*) were excluded from further analysis. For the measures LDB-*dark*, EPM-*eca*, NSF-*lat*, STC-*rrs* we reversed the sign of the effect size, because a decrease in behaviour manifestation is expected as a result of treatment. Our data pool was subset by test measure and a meta-regression model was fitted for every subset.

rma (yi, vi, mods=~factor (compound)−1, random=list(~1|study/observation, ~1|strain,~1|administration mode)

Standardised mean differences (Hedges’ g) were tested with the modifier ‘compound’ (anxiolytic compounds) against the null hypothesis of the estimated effect size for each compound group equalling zero. Publication, strain, and administration mode (namely, whether a compound was administered acutely or chronically) were added as random effects. Additionally, ‘observation’ was added as random effect to account for control groups being used for multiple comparisons (i.e. within the same study and for the same test measure). To assess the overall estimated effect size, independent of anxiolytic compound, the same model syntax was used, excluding the factor modifier. Total and partial I^2, indicating the percentage of sample variation, were used as a measure of heterogeneity, and were calculated using the methods proposed in (Konstantopoulos, 2011).

4.6. Risk of bias

Due to the large sample size, an assessment of quality was made on a

subsample consisting of 180 randomly selected papers. The assessment was done by two independent reviewers (MR, CP), who evaluated 80 different papers each, as well as 20 papers that were reviewed by both investigators, to estimate inter-rater reliability. We used an adapted combination of the CAMARADES study quality checklist and SYRCLE's risk of bias tool, The risk of bias assessment was done by adding specific items, most relevant to our study, from the CAMARADES study quality checklist to the SYRCLE's risk of bias tool(Supplement 1). In particular, we added items such as the statement of compliance with regulatory requirements and the a priori sample size calculation, used in the CAMARADES study quality checklist. On the other hand, CAMARADES' items regarding the use of certain anaesthetics or animals with hypertension or diabetes, do not appear relevant to our study.

Funding statement

This work was supported by the Swiss National Science Foundation, SNF Grant No. 310030-179254 to HW.

Competing interests

The authors declare no competing interests.

Acknowledgements

The authors would like to thank Dr. Cathaljin Leenaars and Dr. Georgia Salanti for their assistance in the data analysis and their valuable feedback on the interpretation of the results.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.neubiorev.2022.104928.

References

Andreatini, Roberto, Bacellar, Leila F.S., 2000. Animal models: trait or state measure? The test-retest reliability of the elevated plus-maze and behavioral despair. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 24 (4), 549–560. https://doi.org/10.1016/S0278-5846(00)00092-0.

Antoniuk, Svitlana, Bijata, Monika, Ponimaskin, Evgeni, Wlodarczyk, Jakub, 2019. Chronic unpredictable mild stress for modeling depression in rodents: meta-analysis of model reliability. Neurosci. Biobehav. Rev. 99, 101–116. https://doi.org/10.1016/j.neubiorev.2018.12.002.

Aron, C., Simon, P., Larousse, C., Boissier, J.R., 1971. Evaluation of a rapid technique for detecting minor tranquilizers. Neuropharmacology 10 (4), 459–469. https://doi.org/10.1016/0028-3908(71)90074-8.

Bailoo, Jeremy D., Reichlin, Thomas, S., Würbel, Hanno, 2014. Refinement of experimental design and conduct in laboratory animal research. ILAR J. 55 (3), 383–391. https://doi.org/10.1093/ilar/ilu037.

Baker, Monya, 2015. Reproducibility crisis: Blame it on the antibodies. Nature 521 (7552), 274–276.

Baker, Monya, 2016. Is there a Reproducibility Crisis? 1,500 scientists lift the lid on reproducibility. In: Nature, 533, pp. 452–454. https://doi.org/10.1126/science.aac4716.

Begley, C.Glenn, Ioannidis, John P.A., 2015. Reproducibility in science: improving the standard for basic and preclinical research. Circ. Res. 116 (1), 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819.

Belzung, Catherine, 1999. Measuring rodent exploratory behavior. Handb. Mol. -Genet. Tech. Brain Behav. Res. 13.

Belzung, Catherine, Griebel, Guy, 2001. Measuring normal and pathological anxiety-like behaviour in mice: a review. Behav. Brain Res. 125 (1–2), 141–149. https://doi.org/10.1016/S0166-4328(01)00291-1.

Benmansour, Saloua, Cecchi, Marco, Morilak, David, A., Gerhardt, Greg, A., Javors, Martin, A., Gould, Georgianna, G., Frazer, Alan, 1999. Effects of chronic antidepressant treatments on serotonin transporter function, density, and mRNA level. J. Neurosci. 19 (23), 10494–10501. https://doi.org/10.1523/JNEUROSCI.19-23-10494.1999.

Bespalov, Anton, Steckler, Thomas, 2021. Pharmacology of anxiety or pharmacology of elevated plus maze. ? Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2020.11.026.

Bodnoff, Shari R., Suranyi-Dacotte, Barbara, Aitken, David H., Quirion, Remi, Meaney, Michael, 1988. The effects of chronic antidepressant treatment in an animal model of anxiety. Psychopharmacology 95, 298–302.

Borsini, Franco, Podhorna, Jana, Marazziti, Donatella, 2002. Do animal models of anxiety predict anxiolytic-like effects of antidepressants? Psychopharmacology 163 (2), 121–141. https://doi.org/10.1007/s00213-002-1155-6.

Bourin, Michel, Petit-Demoulière, Benoit, Dhonnchadha, Brid Nic, Hascöet, Martine, 2007. Animal models of anxiety in mice. Fundam. Clin. Pharmacol. 21 (6), 567–574. https://doi.org/10.1111/j.1472-8206.2007.00526.x.

Bouwknecht, J.Adriaan, Paylor, Richard, 2002. Behavioral and physiological mouse assays for anxiety: a survey in nine mouse strains. Behav. Brain Res. 136 (2), 489–501. https://doi.org/10.1016/S0166-4328(02)00200-0.

Broekkamp, Chris L., Rijk, Huub W., Joly-Gelouin, Danielle, Lloyd, Kenneth L., 1986. Major tranquillizers can be distinguished from minor tranquillizers on the basis of effects on marble burying and swim-induced grooming in mice. Eur. J. Pharmacol. 126 (3), 223–229.

Bystrisky, Alexander, Khalsa, Sahib, S., Cameron, Michael, R., Schiffman, Jason, 2013. Current diagnosis and treatment of anxiety disorders. Pharm. Ther. 38 (1), 30–44.

Cohen, Jacob, 1977. Statistical Power Analysis for the Behavioral Sciences. Academic Press.

Collins, Francis S., Tabak, Lawrence A., 2014. NIH plans to enhance reproducibility. Nature 505 (7485), 612–613.

Contopoulos-Ioannidis, Despina G., Ntzani, Evangelia, E., Ioannidis, John P.A., 2003. Translation of highly promising basic science research into clinical applications. Am. J. Med. 114 (6), 477–484. https://doi.org/10.1016/S0002-9343(03)00013-5.

Costa, J.éssica, Pereira, Oliveira, Guilherme Antônio L., de; Almeida, Antônia Amanda C., de; Islam, Md, Torequl, Sousa, Damião Pergentino, de; Freitas, Rivelilson Mendes, de, 2014. Anxiolytic-like effects of phytol: possible involvement of GABAergic transmission. Brain Res. 1547, 34–42. https://doi.org/10.1016/j.brainres.2013.12.003.

Couzin-Frankel, Jennifer, 2013. When mice mislead. Science 342 (6161), 922-3, 925. DOI: 10.1126/science.342.6161.922.

Crawley, Jacqueline, Goodwin, Frederick K., 1980. Preliminary report of a simple animal behavior model for the anxiolytic effects of benzodiazepines. Pharmacol. Biochem. Behav. 13 167–170.

Crawley, Jacqueline N., 2007. What's wrong with my mouse? Behavioral phenotyping of transgenic and knockout mice, 2nd ed. Wiley-Interscience, Hoboken N.J.

Cryan, John F., Sweeney, Fabian F., 2011. The age of anxiety: role of animal models of anxiolytic action in drug discovery. Br. J. Pharmacol. 164 (4), 1129–1161. https://doi.org/10.1111/j.1476-5381.2011.01362.x.

DeFries, J.C., Hegmann, J.P., Weir, Morton, W., 1966. Open-field behavior in mice: evidence-for a major gene effect mediated by the visual system. Science 154 (3756), 1577–1579.

DeFries, J.C., Hegmann, J.P., Halcomb, R.A., 1974. Response to 20 generations of selection for open-field activity in mice. Behav. Biol. 11, 481–495.

Directive, 2010/63/EU. Additional tools Legislation for the protection of animals used for scientific purposes.

Du Percie Sert, Nathalie, Hurst, Viki, Ahluwalia, Amrita, Alam, Sabina, Avey, Marc T., Baker, Monya, et al., 2020. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. PLoS Biol. 18 (7), e3000410 https://doi.org/10.1371/journal.pbio.3000410.

Ennaceur, A., 2014. Tests of unconditioned anxiety - pitfalls and disappointments. Physiol. Behav. 135, 55–71. https://doi.org/10.1016/j.physbeh.2014.05.032.

Ennaceur, Abdelkader, Chazot, Paul L., 2016. Preclinical animal anxiety research - flaws and prejudices. Pharmacol. Res. Perspect. 4 (2), e00223 https://doi.org/10.1002/prp2.223.

FDA.gov.

File, Sandra E., Hyde, J.R.G., 1979. A test of anxiety that distinguishes between the actions of benzodiazepines and those of other minor tranquilisers and of stimulants. Pharmacol. Biochem. Behav. 11, 65–69.

File, Sandra E., Wardill, Ann G., 1975. The reliability of the hole-board apparatus. Psychopharmacologia 44 (1), 47–51. https://doi.org/10.1007/BF00421183.

Freedman, Leonard P., Cockburn, Iain, M., Simcoe, Timothy S., 2015. The economics of reproducibility in preclinical research. PLoS Biol. 13 (6), e1002165 https://doi.org/10.1371/journal.pbio.1002165.

Gard, Paul R., Haigh, Samantha J., Cambursano, Piero T., Warrington, Claire A., 2001. Strain differences in the anxiolytic effects of losartan in the mouse. Pharmacol. Biochem. Behav. 69, 35–40.

Garner, Joseph P., 2014. The significance of meaning: why do over 90 % of behavioral neuroscience results fail to translate to humans, and what can we do to fix it? ILAR J. 55 (3), 438–456. https://doi.org/10.1093/ilar/ilu047.

Garner, Joseph P., Gaskill, Brianna N., Weber, Elin M., Ahloy-Dallaire, Jamie, Pritchett-Corning, Kathleen R., 2017. Introducing Therioepistemology: the study of how knowledge is gained from animal research. Lab Anim. 46 (4), 103–113. https://doi.org/10.1038/laban.1224.

Geerts, Hugo, 2009. Of mice and men: bridging the translational disconnect in CNS drug discovery. CNS Drugs 23 (11), 915–926.

Genzel, Lisa, Adan, Roger, Berns, Anton, van den Beucken, Jeroen J.J.P., Blokland, Arjan, Boddeke, Erik H.W.G.M., et al., 2020. How the COVID-19 pandemic highlights the necessity of animal research. Curr. Biol. 30 (18), R1014–R1018. https://doi.org/10.1016/j.cub.2020.08.030.

Griebel, Guy, Belzung, Catherine, Perrault, Ghislaine, Sanger, David J., 2000. Differences in anxiety-related behaviours and in sensitivity to diazepam in inbred and outbred strains of mice. Psychopharmacology 148, 164–170.

Hackam, Daniel G., Redelmeier, Danald A., 2006. Translation of research evidence from animals to humans. J. Am. Med. Assoc. 296 (14), 1731–1732.

Hagenbuch, Niels, Feldon, Joram, Yee, Benjamin K., 2006. Use of the elevated plus-maze test with opaque or transparent walls in the detection of mouse strain differences and the anxiolytic effects of diazepam. Behav. Pharmacol. 31–41.

Hall, Calvin S., 1934. Emotional behavior in the rat. I. Defecation and urination as measures of individual differences in emotionality. J. Comp. Psychol. 18 (3), 385–403.
Haller, Jozsef, Alicki, Mano, 2012. Current animal models of anxiety, anxiety disorders, and anxiolytic drugs. Curr. Opin. Psychiatry 25 (1), 59–64. https://doi.org/10.1097/YCO.0b013e32834de34f.
Hånell, Anders, Marklund, Niklas, 2014. Structured evaluation of rodent behavioral tests used in drug discovery research. Front. Behav. Neurosci. 8, 252. https://doi.org/10.3389/fnbeh.2014.00252.
Harro, Jaanus, 2018. Animals, anxiety, and anxiety disorders: how to measure anxiety in rodents and why. Behav. Brain Res. 352, 81–93. https://doi.org/10.1016/j.bbr.2017.10.016.
Hart, Bert A. 't, 2015. Reverse translation of failed treatments can help improving the validity of preclinical animal models. Eur. J. Pharmacol. 759, 14–18. https://doi.org/10.1016/j.ejphar.2015.03.030.
Hascoët, Martine, Bourin, Michel, 1998. A new approach to the light/dark test procedure in mice. Pharmacol. Biochem. Behav. 60 (3), 645–653. https://doi.org/10.1016/S0091-3057(98)00031-8.
Hay, Michael, Thomas, David W., Craighead, John L., Economides, Celia, Rosenthal, Jesse, 2014. Clinical development success rates for investigational drugs. Nat. Biotechnol. 32 (1), 40–51. https://doi.org/10.1038/nbt.2786.
Henderson, Valerie C., Kimmelman, Jonathan, Fergusson, Dean, Grimshaw, Jeremy M., Hackam, Dan G., 2013. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. PLoS Med. 10 (7), e1001489 https://doi.org/10.1371/journal.pmed.1001489.
Howells, David W., Sena, Emily S., Macleod, Malcolm R., 2014. Bringing rigour to translational medicine. Nat. Rev. Neurol. 10 (1), 37–43. https://doi.org/10.1038/nrneurol.2013.232.
Kessler, Ronald C., Berglund, Patricia, Demler, Olga, Jin, Robert, Merikangas, Kathleen R., Walters, Ellen E., 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. Arch. Gen. Psychiatry 62, 593–602.
Kola, Ismail, Landis, John, 2004. Can the pharmaceutical industry reduce attrition rates? Nat. Rev. Drug Discov. 3 (8), 711–715. https://doi.org/10.1038/nrd1470.
Konstantopoulos, Spyros, 2011. Fixed effects and variance components estimation in three-level meta-analysis. Res. Synth. Methods 2 (1), 61–76. https://doi.org/10.1002/jrsm.35.
Kumar, Vijender, Bhat, Zulfiqar Ali, Kumar, Dinesh, 2013. Animal models of anxiety: a comprehensive review. J. Pharmacol. Toxicol. Methods 68 (2), 175–183. https://doi.org/10.1016/j.vascn.2013.05.003.
Langer, Erez, Einat, Haim, Stukalin, Yelena, 2020. Similarities and dissimilarities in the effects of benzodiazepines and specific serotonin reuptake inhibitors (SSRIs) in the defensive marble burying test: a systematic review and meta-analysis. Eur. Neuropsychopharmacol.: J. Eur. Coll. Neuropsychopharmacol. 36, 38–49. https://doi.org/10.1016/j.euroneuro.2020.04.007.
Leenaars, Cathalijn H.C., Kouwenaar, Carien, Stafleu, Frans R., Bleich, André, Ritskes-Hoitinga, Merel, Vries, Rob B.M. de, Meijboom, Franck L.B., 2019. Animal to human translation: a systematic scoping review of reported concordance rates. J. Transl. Med. 17 (1), 223. https://doi.org/10.1186/s12967-019-1976-2.
Leffa, Douglas T., Panzenhagen, Alana, C., Salvi, Artur, A., Bau, Claiton, H.D., Pires, Gabriel, N., Torres, Iraci, L.S., et al., 2019. Systematic review and meta-analysis of the behavioral effects of methylphenidate in the spontaneously hypertensive rat model of attention-deficit/hyperactivity disorder. Neurosci. Biobehav. Rev. 100, 166–179. https://doi.org/10.1016/j.neubiorev.2019.02.019.
Mak, Isabella W.Y., Evaniew, Nathan, Ghert, Michelle, 2014. Lost in translation: animal models and clinical trials in cancer treatment. In: American Journal of Translational Research, 6, pp. 114–118.
Miyakawa, Tsuyoshi, 2020. No raw data, no science: another possible source of the reproducibility crisis. Mol. Brain 13 (1), 24. https://doi.org/10.1186/s13041-020-0552-2.
Moniruzzaman, Md, Mannan, Md. Abdul, Hossen Khan, Md. Farhad, Abir, Ariful Basher, Afroze, Mirola, 2018. The leaves of Crataeva nurvala Buch-Ham. modulate locomotor and anxiety behaviors possibly through GABAergic system. In. BMC Complement. Altern. Med. 18 (1), 283. https://doi.org/10.1186/s12906-018-2338-y.
Montgomery, K.C., 1958. The relation between fear induced by novel stimulation and exploratory behavior. J. Comp. Physiol. PsychoL. 48, 254–260.
O'Collins, Victoria E., Macleod, Malcolm R., Donnan, Geoffrey A., Horky, Laura L., van der Worp, Bart H., Howells, David W., 2006. 1,026 experimental treatments in acute stroke. Ann. Neurol. 59 (3), 467–477. https://doi.org/10.1002/ana.20741.
Ohl, Frauke, Arndt, Saskia S., van der Staay, F.Josef, 2008. Pathological anxiety in animals. Vet. J. (Lond., Engl.: 1997) 175 (1), 18–26. https://doi.org/10.1016/j.tvjl.2006.12.013.
Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H.-U., Jönsson, B., 2012. The economic cost of brain disorders in Europe. Eur. J. Neurol. 19 (1), 155–162. https://doi.org/10.1111/j.1468-1331.2011.03590.x.
Perrin, Steve, 2014. Make Mouse Studies Work. In: Nature, 507, pp. 423–425. https://doi.org/10.1016/j.brainres.2013.10.013.
Pires, Gabriel Natan, Bezerra, Andréia Gomes, Tufik, Sergio, Andersen, Monica Levy, 2016. Effects of experimental sleep deprivation on anxiety-like behavior in animal research: systematic review and meta-analysis. Neurosci. Biobehav. Rev. 68, 575–589. https://doi.org/10.1016/j.neubiorev.2016.06.028.
Pound, Pandora, Bracken, Michael B., 2014. Is animal research sufficiently evidence based to be a cornerstone of biomedical research? BMJ (Clin. Res. Ed.) 348, g3387. https://doi.org/10.1136/bmj.g3387.
Prut, Laetitia, Belzung, Catherine, 2003. The open field as a paradigm to measure the effects of drugs on anxiety-like behaviors: a review. Eur. J. Pharmacol. 463 (1–3), 3–33. https://doi.org/10.1016/S0014-2999(03)01272-X.
R Core Team (2020): R: A Language and Environment for Statistical Computing. Vienna, Austria. Available online at ⟨https://www.R-project.org/⟩.
Ravindran, Lakshmi N., Stein, Murray B., 2010. The pharmacologic treatment of anxiety disorders: a review of progress. J. Clin. Psychiatry 71 (7), 839–854.
Rodgers, R.J., Cao, B.-J., Dalvi, A., Holmes, A., 1997. Animal models of anxiety: an ethological perspective. Braz. J. Med Biol. Res 30 (3), 289–304. https://doi.org/10.1590/S0100-879X1997000300002.
Roth, Kevin A., Cox, Audra E., 2015. Science isn't science if it isn't reproducible. Am. J. Pathol. 185 (1), 2–3. https://doi.org/10.1016/j.ajpath.2014.11.001.
Russell, William Moy Stratton, Burch, Rex Leonard, 1959. The Principles of Humane Experimental Technique. Methuen.
Santana, Lorena C.L.R., Brito, Maria, R.M., Oliveira, George, L.S., Citó, Antônia, M.G.L., Alves, Clayton, Q., David, Juceni, P., et al., 2014. Mikania glomerata: phytochemical, pharmacological, and neurochemical study. Evid. -Based Complement. Altern. Med.: eCAM 2014, 710410. https://doi.org/10.1155/2014/710410.
Shakhnovich, Valentina, 2018. It's time to reverse our thinking: the reverse translation research paradigm. Clin. Transl. Sci. 11 (2), 98–99. https://doi.org/10.1111/cts.12538.
Shepherd, Jon K., Grewal, Savraj S., Fletcher, Allan, Bill, David J., Dourish, Colin T., 1994. Behavioural and pharmacological characterisation of the elevated "zero-maze" as an animal model of anxiety. Psychopharmacology 116 (1), 56–64. https://doi.org/10.1007/BF02244871.
Smith, Adrian J., Lilley, Elliot, 2019. The role of the three Rs in improving the planning and reproducibility of animal experiments. Anim.: Open Access J. MDPI 9 (11). https://doi.org/10.3390/ani9110975.
Smith, Adrian J., Clutton, R.Eddie, Lilley, Elliot, Hansen, Kristine E.Aa, Brattelid, Trond, 2018. PREPARE: guidelines for planning animal research and testing. Lab. Anim. 52 (2), 135–141. https://doi.org/10.1177/0023677217724823.
Steimer, Thierry, 2011. Animal models of anxiety disorders in rats and mice: some conceptual issues. Dialog-. Clin. Neurosci. 13 (4), 495–506.
Thiébot, Marie-H.élène, Soubrié, Philippe, Simon, Pierre, 1985. Is delay of reward mediated by shock-avoidance behavior a critical targer for anti-punishment effects of diazepam in rats? Psychopharmacology 473–479.
van der Staay, F.Josef, Arndt, Saskia S., Nordquist, Rebecca E., 2009. Evaluation of animal models of neurobehavioral disorders. Behav. Brain Funct.: BBF 5, 11. https://doi.org/10.1186/1744-9081-5-11.
van der Worp, H.Bart, Howells, David W., Sena, Emily S., Porritt, Michelle J., Rewell, Sarah, O'Collins, Victoria, Macleod, Malcolm R., 2010. Can animal models of disease reliably inform human studies? PLoS Med. 7 (3), e1000245 https://doi.org/10.1371/journal.pmed.1000245.
Viechtbauer, Wolfgang, 2010. Conducting meta-analyses in R with the metafor Package. J. Stat. Softw. 36, 3.
Voelkl, Bernhard, Vogt, Lucile, Sena, Emily, S., Würbel, Hanno, 2018. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS Biol. 16 (2), e2003693 https://doi.org/10.1371/journal.pbio.2003693.
Voelkl, Bernhard, Altman, Naomi S., Forsman, Anders, Forstmeier, Wolfgang, Gurevitch, Jessica, Jaric, Ivana, et al., 2020. Reproducibility of animal research in light of biological variation. Nat. Rev. Neurosci. 21 (7), 384–393. https://doi.org/10.1038/s41583-020-0313-3.
Vos, Theo, Allen, Christine, Arora, Megha, et al., 2016. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 388 (10053), 1545–1602. https://doi.org/10.1016/S0140-6736(16)31678-6.
Wahlsten, Douglas, 2001. Standardizing tests of mouse behavior: reasons, recommendations, and reality. Physiol. Behav. 73 (5), 695–704. https://doi.org/10.1016/S0031-9384(01)00527-3.
Wahlsten, Douglas, Rustay, Nathan R., Metten, Pamela, Crabbe, John C., 2003. In search of a better mouse test. Trends Neurosci. 26 (3), 132–136. https://doi.org/10.1016/S0166-2236(03)00033-X.
Walsh, Roger N., Cummins, Robert A., 1976. The open-field test: a critical review. Psychol. Bull. 83 (3), 482–504.
Wicherts, Jelte M., Veldkamp, Coosje, L.S., Augusteijn, Hilde, E.M., Bakker, Marjan, van Aert, Robbie C.M., van Assen, Marcel A.L.M., 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. Front. Psychol. 7, 1832. https://doi.org/10.3389/fpsyg.2016.01832.
Willner, P., 1984. The validity of animal models of depression. Psychopharmacology 83, 1–16.
Würbel, H., 2000. Behaviour and the standardization fallacy. Nat. Genet. 26, 262–263.
Würbel, Hanno, 2017. More than 3Rs: the importance of scientific validity for harm-benefit analysis of animal research. Lab Anim. 46 (4), 164–166. https://doi.org/10.1038/laban.1220.