From replicability to generalizability

Arroyo Araujo, Maria

DOI:
10.33612/diss.325014460

FROM REPLICABILITY TO GENERALIZABILITY

How research practice can shape scientific results

María Arroyo Araujo

ISBN 978-94-6361-797-0

Title:	From replicability to generalizability
Author:	María Arroyo Araujo
Cover & Design:	Claudia Arroyo Araujo
Lay-out:	Optima Grafische Communicatie (www.ogc.nl)
Printed by:	Optima Grafische Communicatie (www.ogc.nl)

university of groningen

From replicability to generalizability

How research practice can shape scientific results

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Tuesday 17th January 2023 at 12:45 hours

by

María Arroyo Araujo

born on 3 October 1990
in Mexico City, Mexico

Supervisors
Prof. M.J.H. Kas
Prof. R. Havekes

Co-supervisor
Dr. P. Meerlo

Assessment Committee
Prof. G. Van Dijk
Prof. J.D.A. Olivier
Prof. L.M. Bouter

Para mi mamá,

Por ser mi mayor fuente de inspiración y fortaleza.

To my mother,

For being my greatest source of inspiration and strength.

The studies described in this thesis were performed at the Groningen Institute for Evolutionary Life Sciences (GELIFES) and the Behavioral and Cognitive Neurosciences (BCN) Research School of the University of Groningen, Groningen, the Netherlands.

This project has received funding from the European Autism Interventions - A Multicenter Study for Developing New Medications (EU-AIMS) project, which receives support from the Innovative Medicines Initiative Joint Undertaking under Grant agreement number 115300, composed of financial contributions from the European Union's Seventh Framework Programme (FP7/2007-2013), the European Federation of Pharmaceutical Industries and Associations companies-and from Autism Speaks.

The Innovative Medicines Initiative 2 Joint Undertaking under (grant number 777364, 2017); this Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA for the European Quality in Preclinical Data (EQIPD) consortium.

The ZonMW MKMD subsidy (grant number 114024154, 2020) supported the meta-analysis chapter of this thesis

CONTENTS

General introduction

María Arroyo-Araujo

Groningen Institute of Evolutionary Life Sciences (GELIFES), University of Groningen (Groningen, the Netherlands)

REPRODUCIBILITY AND REPLICABILITY OF RESULTS

The development of new therapeutic targets relies heavily on the results of preclinical research. For this reason, it is of the utmost importance that preclinical findings are reproducible and replicable. In brief, reproducible results say about the feasibility to obtain consistent results using the same input data, methods, code, analysis (*i.e.,* computational reproducibility). Thus, reproducibility is closely linked to transparency and does not have a say about the correctness of the computation (*e.g.,* if there is an error in the experimental design and the study is replicated, the same erroneous result will be replicated). On the other hand, replicability means obtaining consistent results after collecting new data by using comparable methodologies (1,2). The relevance of replicable findings in preclinical studies lays on maximizing the potential of these findings towards the development of therapeutic strategies for the target (clinical) population. Specifically, replicability of results from scientific research has served as a way to operationalize truth as it suggests that the phenomenon under examination can be detached from the specific circumstances at which it was originally assessed (2,3). Unfortunately, over the past decades there have been numerous accounts of poor scientific replicability both across and within labs, certainly preclinical research and rodent phenotyping studies are not the exception (4,5). For instance, the landmark multi-center study by Crabbe et al., (1999) (6) showed that the variability across laboratories was larger than within laboratories following protocol standardization and harmonization across sites. The consequences of this 'replicability crisis' impact both the scientific community as well as the general public. For example, researchers may be unintentionally misled by inconclusive and/or inaccurate findings steering research towards slow, non-efficient, treatment development for clinical trials. Other costs as a result of poor replicability of results include the waste of financial and other resources, ethical concerns that come with the use of animals for inconclusive/uninformative research, as well as the delay in development of new therapeutic treatments. Thus, there is an urgent need to identify the underlying causes of irreplicable scientific findings, thereby reduce variability across and within laboratories, and ultimately improve the scientific value of preclinical studies for the development of novel therapeutic strategies that could benefit patients and their families.

Sources and countermeasures

The possible sources for irreplicable results are numerous and differ on their potential to help gaining knowledge. According to this classification, irreplicable results that are helpful to gain knowledge are consequence of studying complex systems with imperfect knowledge and tools and they represent a normal part of the scientific process rather than mistakes. In contrast, irreplicable results that are not helpful for gaining knowledge

come from shortcomings in the design, performance and reporting of studies; these can be honest mistakes or deliberate misconduct (1).

When it comes to replicability of results, it is necessary to consider the study's internal validity: a study is said to have internal/causal validity when one can assure that the outcome obtained was caused by the experimental manipulation and not by any other source of variability (7). To ensure this causal relationship, experimental designs should account for unknown sources of variability (*i.e.,* noise) that could influence the effect of the experimental manipulation. This can be achieved by research practices that minimize the risk of bias (*e.g.,* blinding of groups/treatments, randomization of subjects, etc.) and thus, prevent possible confounding. Hence, results from studies with internal validity are more likely to be accurate and replicable.

Other sources of irreplicable results and/or low interval validity that are classified as unhelpful to gain knowledge are: publication bias, underpowered studies, p-hacking, and p-HARking (Hypothesis After Results). Such suboptimal research practices are indeed some of the best known sources of irreplicable results (4) and have a tight link to research integrity. According to the national survey on research integrity (NSRI) performed to Dutch researchers across fields and academic ranks, 50% of responders have engaged in at least 1 of 11 questionable research practices (QRPs) surveyed (8). In addition to QRPs, falsification, fabrication and plagiarism (FFP) also affect replicability; these sources are associated to researcher integrity. Such researcher misconduct is more rare in frequency than QRPs and so its impact on the literature is smaller (9); nevertheless, 4% of responders engaged in data fabrications at least one time over the previous three years. The results of the NSRI indicate higher prevalence of QRPs than previously reported elsewhere (9,10), indicating that there is a need for a change in the current research culture. As a matter of fact, the authors of the NSRI also explored possible explanatory factors to incur in QRPs and misconduct. Publication pressure was identified as the main driver to engage in QRPs; this pressure is likely to drive researcher towards 'cherry picking' their results towards positive findings which are more easily published than negative ones. Selective reporting biases the body of knowledge contributing heavily to the replicability crisis.

In terms of the academic rank, being a PhD student or junior researcher increased the likelihood of engaging in any QRPs, while scientific norm subscription decreased the odds of QRPs and FF. Together, these findings suggest that better training and mentoring from the supervisors' side is needed to improve the scientific performance of young researchers.

Certainly, there is an urgent need to foster responsible research practices among researchers; besides the integrity of the research and researcher, it is also crucial to improve how research in performed, thus, the work in this thesis will focus on the impact that experimental designs and reporting practices have on the variability of results between studies/laboratories.

It is important to keep in mind that although results reproducibility and replicability are a way to confirm scientific findings, this does not necessarily mean that results can be extrapolated to different contexts (*i.e.,* results generalizability) (11). For example, if a study is successfully replicated across independent samples of male mice, these results may not be informative for female mice. Therefore, in order for results to be generalizable and, thus, likely replicable, they must be sufficiently robust, as will be further discussed in the next section.

Robustness of data

In order to replicate results, the outcomes should be consistent albeit the changes implied when re-running the same experiment in somewhat different times/conditions/populations. Data that is resilient to experimental variation (e.g., environmental and genetic variability) will be more likely to generalize to other contexts (i.e., results will have external validity); thus, they will more likely be reproduced (12).

Experimental design

One way in which the robustness of data can be established is through the experimental design as this sets the boundaries for the contexts to which the results may be able to generalize. Currently, best scientific practices advocate for standardizing the animal subjects and their environment by keeping their properties constant overall (12). While standardizing environmental factors is believed to reduce background noise, when taken to an extreme it will provide results that are only informative for, and replicable under the same circumstances in which they were obtained (*i.e.,* idiosyncratic results) (13–16). This is known as the 'standardization fallacy' (11). One of the main setbacks of rigorous standardization of experimental subjects and their environments is that it fails to incorporate the changes in the expression of a phenotype in response to the environmental influences (12). This makes results more likely to be replicated in the same/similar contexts/settings/times than in novel ones (*i.e.,* less robust). Therefore, experimental designs that incorporate diversity of experimental conditions will result in data that is more likely resilient to diverse contexts. However empirical evidence is needed to support this claim.

Representativeness of study samples

Another way to address robustness of data is to increase the representativeness of the study sample. Representativeness of a study population in the context of preclinical research implies that the study population incorporates the biological variability of the population of interest (3). However, rigorous standardization practices aim to reduce the variability of subjects within a study, which would potentially draw the experimental sample further from the target population. In other words, this approach may decrease the representativeness of the study sample and the likelihood of replicating the results under slightly different conditions due to compromised generalizability (2,11).

A way to improve the representativeness of study populations is by diversifying the rearing/husbandry conditions and/or population characteristics (*e.g.,* age, sex, genetic background) within a study; in other words, creating a more heterogeneous population. In this way the between study variability would decrease as each individual study accounts for the unavoidable differences of phenotypic expression between studies/labs (15,17–20).

Altogether, when aiming to produce replicable results that are informative to a target population, one must minimize the risk of bias and ensure the robustness of data. This can be achieved by practices such as blinding and randomization, and likely by diversifying the experimental conditions and population.

TOWARDS RESEARCH (REPLICABILITY AND REPRODUCIBILITY) IMPROVEMENT

As stated in the section above, it has become clear that the way preclinical experiments are usually planned and conducted should be modified if we intend to improve data quality and data interpretation. Specifically, there is room for improvement in the transparency of reporting of studies, particularly, at the level of the experimental design and the representativeness of the study samples. Towards this end, there have been numerous efforts contributing to improve the replicability and reproducibility of preclinical results by promoting rigorous research practices and informative statistical methods such as the ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines to improve the quality of reporting in animal research (21), and the development of initiatives like SYRCLE (Systematic Review Center for Laboratory animal Experimentation) to guide and provide tools to improve the appraisal of systematic reviews and meta-analyses of preclinical data (22). Another initiative is the most recent creation of the open-access Platform for the Exchange of Experimental Research Standards (PEERS) (23) that aims

to provide researchers with a guide on the factors that are most relevant for their experimental design. Overall, the improvement of preclinical studies aims to produce preclinical results that are more accurate, meaningful and informative for translational researchers. For this reason, it is sensible to explore the possible experimental factors affecting data variability in preclinical studies. Certainly, in this way, research practices and study protocols may be adapted accordingly to enhance accuracy, reproducibility and replicability of results across sites.

AIMS AND THESIS OUTLINE

This thesis discusses and tests different perspectives from which data quality and replicability of preclinical results are affected by research practices, and possible ways how to counteract this.

In **Chapter 2,** three different labs part of the EU-AIMS (European Autism Interventions) consortium conducted the same pharmacological experiment in the Shank2-Knockout (KO) rat autism model. The approach taken was to align the apparatus, protocol and data analysis across sites to evaluate their effect on the variability of outcomes between laboratories. Towards this end, the different behavioral outcomes as well as the impact of a pharmacological manipulation were compared across sites. The work described in **Chapter 3** is part of the EQIPD (European Quality in Preclinical Data) consortium aimed to promote research practices that enhance data quality in preclinical studies. This chapter summarizes a three-stage study that evaluated the effects of experimental protocols that differed in the degree of standardization within-laboratory and harmonization across seven labs from academia and industry. The aim of this study was to evaluate impact of protocol harmonization in the variability of results between laboratories.

In **Chapter 4** we investigate the replicability of the behavioral phenotype of the Fragile-X mental retardation mouse model (Fmr1-KO) through a systematic review and meta-analysis. This analysis includes a report on transparency of reporting in light of data quality and replicability of results. Based on the results of this meta-analysis, a study was carried out to assess the behavioral phenotype of the Fmr1-KO. In addition, we explored the possible influence of specific environmental factors that are in fact often not reported in scientific literature. We assessed whether these factors can act as a source of variability thereby potentially contributing to poor replicability. The results from this behavioral study is presented in **Chapter 5**.

As mentioned above, part of the work described in this thesis was carried out as part of the EQIPD consortium. Related to this, in **Chapter 6** we describe the implementation of the quality system (QS), developed by EQIPD members, in an academic lab setting. This chapter exemplifies how the use of this tool can promote rigorous research practices to boost preclinical data quality in academia and industry. More detailed information regarding the EQIPD-QS can be found in **Appendix 1** of this thesis. Finally, **Chapter 7** discusses the overall findings of this thesis, provides conclusions and future perspectives on data quality and replicability of preclinical data.

REFERENCES

1. National Academies of Sciences, Engineering, and Medicine. Reproducibility and Replicability in Science. [Internet]. Washington, DC: The National Academies Press.; 2019. Available from: https://doi.org/10.17226/25303.
2. Kafkafi N, Agassi J, Chesler EJ, Crabbe JC, Crusio WE, Eilam D, et al. Reproducibility and replicability of rodent phenotyping in preclinical studies. Neurosci Biobehav Rev. 2018 Apr;87:218–32.
3. Kukull WA, Ganguli M. Generalizability: the trees, the forest, and the low-hanging fruit. Neurology. 2012 Jun 5;78(23):1886–91.
4. Bishop D. Rein in the four horsemen of irreproducibility. Nature. 2019 Apr;568(7753):435–435.
5. Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, et al. Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. PLOS ONE. 2009 Nov 30;4(11):e7824.
6. Crabbe JC, Wahlsten D, Dudek BC. Genetics of mouse behavior: interactions with laboratory environment. Science. 1999 Jun 4;284(5420):1670–2.
7. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull. 1959 Mar;56(2):81–105.
8. Gopalakrishna G, Riet G ter, Vink G, Stoop I, Wicherts JM, Bouter LM. Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. PLOS ONE. 2022 Feb 16;17(2):e0263023.
9. Fanelli D. Is science really facing a reproducibility crisis, and do we need it to? Proc Natl Acad Sci. 2018 Mar 13;115(11):2628–31.
10. Xie Y, Wang K, Kong Y. Prevalence of Research Misconduct and Questionable Research Practices: A Systematic Review and Meta-Analysis. Sci Eng Ethics. 2021 Jun 29;27(4):41.
11. Würbel H. Behaviour and the standardization fallacy. Nat Genet. 2000 Nov;26(3):263.
12. Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, Jaric I, et al. Reproducibility of animal research in light of biological variation. Nat Rev Neurosci. 2020 Jul;21(7):384–93.
13. Paylor R. Questioning standardization in science. Nat Methods. 2009 Apr;6(4):253–4.
14. Richter SH, Garner JP, Würbel H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nat Methods. 2009 Apr;6(4):257–61.
15. Richter SH, Garner JP, Auer C, Kunert J, Würbel H. Systematic variation improves reproducibility of animal experiments. Nat Methods. 2010 Mar;7(3):167–8.
16. Wahlsten D, Metten P, Phillips TJ, Boehm SL, Burkhart-Kasch S, Dorow J, et al. Different data from different labs: lessons from studies of gene-environment interaction. J Neurobiol. 2003 Jan;54(1):283–311.
17. Bodden C, Kortzfleisch VT von, Karwinkel F, Kaiser S, Sachser N, Richter SH. Heterogenising study samples across testing time improves reproducibility of behavioural data. Sci Rep. 2019 Jun 3;9(1):8247.
18. Farrar BG, Voudouris K, Clayton N. Replications, Comparisons, Sampling and the Problem of Representativeness in Animal Cognition Research [Internet]. PsyArXiv; 2020 Aug [cited 2022 Jan 3]. Available from: https://osf.io/2vt4k
19. Voelkl B, Vogt L, Sena ES, Würbel H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLOS Biol. 2018 Feb 22;16(2):e2003693.
20. von Kortzfleisch VT, Karp NA, Palme R, Kaiser S, Sachser N, Richter SH. Improving reproducibility in animal research by splitting the study population into several 'mini-experiments.' Sci Rep. 2020 Oct 6;10(1):16579.

21. Sert NP du, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, et al. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLOS Biol. 2020 Jul 14;18(7):e3000411.
22. Hooijmans CR, Draper D, Ergün M, Scheffer GJ. The effect of analgesics on stimulus evoked pain-like behaviour in animal models for chemotherapy induced peripheral neuropathy- a meta-analysis. Sci Rep. 2019 Nov 26;9(1):17549.
23. Sil A, Bespalov A, Dalla C, Ferland-Beckham C, Herremans A, Karantzalos K, et al. PEERS — An Open Science "Platform for the Exchange of Experimental Research Standards" in Biomedicine. Front Behav Neurosci [Internet]. 2021 [cited 2022 Feb 12];15. Available from: https://www.frontiersin.org/article/10.3389/fnbeh.2021.755812

2

Reproducibility via coordinated standardization: a multi-center study in a Shank2 genetic rat model for Autism Spectrum Disorders

María Arroyo-Araujo[1]*, Radka Graf[2]*, Martine Maco[3]*, Elsbeth van Dam[4], Esther Schenker[5], Wilhelmus Drinkenburg[6], Bazstijn Koopmans[7], Sietse F. de Boer[1], Michaela Cullum-Doyle[2], Lucas P.J.J. Noldus[4], Maarten Loos[7], Wil van Dommelen[4], Will Spooren[2], Barbara Biemans[3#], Derek L. Buhl[2#], and Martien J. Kas[1,8#]

* Shared first authorship
# Shared last authorship
[1]Groningen Institute for Evolutionary Life Sciences, University of Groningen (Groningen, the Netherlands)
[2]Neuroscience and Pain Research Unit, Pfizer Inc. (Cambridge, MA, USA)
[3]Roche Innovation Center Basel, (Basel, Switzerland)
[4]Noldus Information Technology BV (Wageningen, the Netherlands)
[5]Institut de Recherches Servier (Croissy-sur-Seine, France)
[6]Janssen Research & Development, Janssen Pharmaceutica NV (Beerse, Belgium)
[7]Sylics (Synaptologics BV, Amsterdam, the Netherlands)
[8]Department of Translational Neuroscience, University Medical Center Utrecht (Utrecht, the Netherlands)

Published in Scientific Reports (2019) 9:11602.

ABSTRACT

Inconsistent findings between laboratories are hampering scientific progress and are of increasing public concern. Differences in laboratory environment is a known factor contributing to poor reproducibility of findings between research sites and well-controlled multisite efforts are an important next step to identify the relevant factors needed to reduce variation in study outcome between laboratories. Through harmonization of apparatus, test protocol, and aligned and non-aligned environmental variables, the present study shows that behavioral pharmacological responses in *Shank2* knockout (KO) rats, a model for autism spectrum disorders, were highly replicable across three research centers. All three sites reliably observed a hyperactive and repetitive behavioral phenotype in KO rats compared to their wild-type littermates as well as a dose-dependent phenotype attenuation following acute injections of a selective mGluR1 antagonist. These results show that reproducibility in preclinical studies can be obtained and emphasizes the need for high quality and rigorous methodologies in scientific research. Considering the observed external validity, the present study also reveals important implications for the treatment efficacy of mGluR1 antagonism for a translational phenotype related to a major risk gene for Autism Spectrum Disorders.

INTRODUCTION

The alarmingly high estimate of failure (50-80%) to replicate findings in preclinical studies is a prevalent issue of great scientific and public concern that needs to be addressed [2, 10, 15]. While the lack of reproducibility of scientific findings has gained significant attention, thus far not many attempts and strategies have been implemented to tackle this challenging situation. Given that the ability to replicate empirical findings is a prerequisite of experimental science, deficient reproducibility hinders scientific credibility and progress. For biomedical animal research in particular, poor reproducibility questions the benefit of research in the ethical analysis of animal experiments [28], prevents pharmacotherapeutic development, and results in great monetary loss [8, 13]. The inability to replicate scientific findings points toward systematic inefficiencies in the way studies are planned, executed, analyzed, and reported. Although drivers of data variability across different research sites are not well understood, the use of different animal strains, housing, husbandry, and testing environments, and/or different lab standard operating procedures (SOPs) are generally considered critical factors [9, 23]. Therefore, rigorous genetic (animals) and environmental (housing, husbandry, testing procedures) standardization have been advocated as good laboratory practice to reduce variation in experimental results [29]. However, excessive standardization results in more homogeneous study populations, which in turn generates spurious results as they are representative to the specific standardized conditions under which the data was obtained, thereby hampering replicability [18, 23, 24, 26].

Gene-environment interactions can considerably affect animal behavior. As laboratories differ in many factors (personnel, odors, noise, microbiota, etc.) and the intrinsic variability of the animals assessed is high (i.e. genetic variation from different vendors), the variation of phenotypes between laboratories, even in the same genetic strain of animals, is generally much larger than the variation within laboratories as clearly shown by the landmark multi-laboratory study of [5]. Then, contrary to common belief, excessive standardization, understood as controlled environmental enrichment that decreases biologically meaningful variation, doesn't contribute to highly reproducible results.

To improve reproducibility of preclinical studies and maximize the chances of discovering meaningful treatment effects or fundamental biological principles, several suggestions have been proposed [3, 13, 21]; in particular, to take into account unavoidable between-laboratory variations. For this, the use of multi-laboratory study designs has been advocated as a valuable approach to evaluate the influence of heterogenization between different laboratory settings on data variability [23]. Using the genetically modified *Shank2* knockout (KO) rat model of autism spectrum disorders reported to exhibit

autistic-like hyperactive and repetitive behavioral phenotype [17], the primary objective of the current study was to investigate whether these previously reported results could be reproduced and replicated across three study sites by following the same experimental protocol for behavioral evaluation with automated video scoring analysis and drug testing. To reduce the impact of environmental factors that typically differ greatly between laboratories and are difficult to control, an identical test setup, i.e. a PhenoTyper® cage and EthoVision XT 12 video tracking software (Noldus Information Technology BV; Wageningen, The Netherlands) was used at all sites. *Shank2* KO rats were placed in this novel environment in an attempt to reproduce the previously observed hyperactive and repetitive behavioral phenotype of these animals [17] that recapitulate the characteristic behavioral abnormalities of autism spectrum disorder (ASD) in humans. In addition, to confirm the normalization after pharmacological treatment with the metabotropic Glutamate Receptor 1 (mGluR1) antagonist JNJ16259685, we included a dose-response of the drug treatment to strengthen interpretation for the effect. Finally, as a secondary objective, a comparison between 2 behavioral scoring methods to evaluate the phenotype (i.e., automated versus manual scoring) was performed using the same recorded videos.

MATERIALS AND METHODS

General study design

Based on the demonstration of hyperactivity and repetitive circling behaviors observed in the *Shank2* KO rat model [17], a cross-site study focusing on these behaviors was initiated. Specifically, we aimed to assess the behavioral phenotype, as well as the pharmacological effects in both *Shank*2 KO and littermate controls (WT) using automated video scoring. A phenotypic assessment was carried out quasi-simultaneously (i.e. during the same month) in three different research facilities. Although the aim of this study was to explore the reproducibility of the results, it was not intended to fully reproduce the original methodology; standardized phenotyping equipment was used, and small changes were made to the protocol for this study.

To optimize the chance of successful replication, the protocol entailed controlling several aspects of the study design from animal provider and shipment, to details of experimental procedures. In addition, some other factors were not harmonized across sites presumably increasing the robustness of the study results.

To enable consistency in the environmental aspects of the behavioral assays and automated scoring methods between sites, and in addition to the PhenoTyper cham-

bers provided, the operational definition of the behavioral categories to score were aligned across sites to pursue consistency in the manual scoring (e.g. what represents a turn). Janssen Pharmaceutica NV (Beerse, Belgium) provided the mGluR1 antagonist JNJ16259685 [14, 21] to all sites to eliminate variability in pharmacological outcomes due to inconsistencies of the chemical batch (e.g. differences in purity).

Laboratories

The experiment was conducted at the Groningen Institute for Evolutionary Life Sciences (GELIFES) of the University of Groningen (RUG, Groningen, The Netherlands), the Neuroscience and Pain Research Unit of Pfizer Inc. (Cambridge, MA, USA), and at the Roche Innovation Center Basel (Basel, Switzerland).

Study design

The experiment was carried out during four consecutive weeks of four testing days per week (Monday to Thursday), always starting 4 hours after onset of light. A full crossover design was followed, so that each subject received each dose once with a one-week wash-out period before the next dosing and testing; for this, on each day of the week 3 WT and 3 KO were tested, so that at the end of the week all subjects (12WT/12KO) were tested once with one of the four treatments (including vehicle).

On a given testing day, each subject was weighed and given a single-dose injection of either vehicle (saline) or JNJ16259685 (0.02, 0.04, or 0.63 mg/kg in 5 ml/kg volume), administered subcutaneously in the flank. The dosing order was alternated between genotypes and counterbalanced across days. The treatment conditions were randomized throughout the experiment with a Latin-square design. Thirty minutes after dosing, the subject was placed in a PhenoTyper chamber and video-recorded for 30 minutes after which the chamber was cleaned with alcohol wipes. The behavior was scored after the experiment from the video images using EthoVision XT 12 for the automated scoring or The Observer XT 13 for the manual scoring. For the latter, the observer was blinded to genotype and treatment.

Animals

36 Sprague-Dawley male rats carrying a targeted deletion of the *Shank2* gene (KO) and 36 male WT rats matched for age were generated as described by Modi et al., 2018. A batch of 12 KO and 12 WT rats was shipped from Charles River Laboratories (Wilmington, MA, USA) to each of the three sites involved. Animals had at least ~4 weeks of habituation to their housing facility and were around 3 months old at the start of the experiment. Animals had *ad libitum* access to food and water with a similar 12:12 light-dark cycle at all three sites.

All animal procedures were carried out following the regulations of the Directive 2010/63/EU and in accordance with the recommendations of the Guide for the Care and Use of Laboratory Animals. The protocol was approved by the Pfizer Institutional Animal Care and Use Committee, the Basel Cantonal Animal Protection Committee adhering to Swiss federal legislation, and the University of Groningen's Animal Welfare Body in accordance with the Central Committee for Animal Experiments.

Drug

The test compound, JNJ16259685-AAA (3,4-dihydro-2H-pyrano[2,3]b-quinolin-7-yl) (cis-4-methoxycyclohexyl) methanone, is a brain penetrant selective mGluR1 antagonist with an affinity of 0.34 nM (Ki value) for rat mGlu1 receptor, which potently and completely inhibits glutamate-induced increases in intracellular Ca2+ concentrations with an IC50 value of 3.24nM (Lavreysen et al. 2004). The compound was synthesized at Janssen Pharmaceutica NV and centrally shipped to the participating labs, by the Compound Logistics & Formulation unit at Janssen Pharmaceutica NV. All sites used compound from the same chemical batch.

JNJ16259685-AAA was dissolved in saline (i.e., H2O+HCL+NaCL to reach a pH of 7.4) and serial dilutions were made for the different doses.

Equipment and software

A PhenoTyper 4500 behavioral assessment chamber (Noldus Information Technology BV) was shipped to each of the three sites, to ascertain standardization of behavioral recording and analysis. The PhenoTyper 4500 chamber includes a black square arena (floor area 45x45cm), 4 matted walls with ventilation holes at the top (66cm tall). The top unit serves as a lid from which only the infrared sensitive camera (30 fps at 640x480 resolution using NTSC format) and the 3 arrays of dimmable infrared LED lights were used. Automated scoring of rat's behavior was done using EthoVision XT 12 video tracking software, including the Rat Behavior Recognition module (Noldus Information Technologies BV) which allowed a repeatable, objective, and consistent analysis of the 30 minutes video. Details of the acquisition settings are listed in Table 1 of the Supplementary material. The video files that were run offline through EthoVision XT 12, were scored using The Observer XT 13 software by a blinded scorer at each of the three sites; for this, only the second 10-minute bin was analyzed.

Behavioral Readouts (automated and manual)

Following a predetermined set of criteria (Tables 1 and 2), behavior was analyzed using two methods of scoring. The predetermined criteria were discussed and agreed upon in detail by the three sites.

Automated scoring

Table 1 lists the behaviors and their definition as recognized by EthoVision XT 12. The automated scores of these behaviors are based on the entire 30 minutes of the experimental session.

Table 1. Description of the detection criteria for the automated tracking using EthoVision XT 12

Variable	Description
Circling	Rotation based on direction from tail-base to center-point, clockwise, count every 0.75 rotation, threshold 30 degrees (frequency)
Circling 2	Rotation based on direction from tail-base to center-point, counterclockwise, count every 0.75 rotation, threshold 30 degrees (frequency)
Rearing Supported	Probability greater than 50%, excludes instances shorter than 0.50s (frequency, cumulative duration)
Rearing Unsupported	Probability greater than 80%, excludes instances shorter than 0.50s (frequency, cumulative duration)
Movement	Averaging interval of 1 sample, start velocity 5.00 cm/s, stop velocity 1.00 cm/s based on the body center-point (mean, cumulative duration)
Walking	Probability greater than 10%, exclude instances shorter than 0.00s (frequency, cumulative duration)

Manual scoring

The manual scoring of behaviors was done by a trained observer blind to genotype and treatment, using The Observer XT 13. Table 2 depicts the behavioral definitions on which the scoring was based. This scoring was only carried out for the second 10-minute bin of the experimental session (i.e. from minute 11 to minute 20).

Table 2. Description of the behavioral definitions used for the manual scoring with The Observer XT 13.

Behavior	How to manually score rat behavior in The Observer XT
Circling (total time)	Start scoring when the rat moves in rapid circles in the same direction lacking apparent goal or function, do not stop until the rat finishes circling. Don't score each full rotation separately, score it as a bout
Circling (frequency)	Score each event when the rat turns in a full circular motion (as reference, the nose has to travel at least 270 degrees)
Rearing Unsupported/ Supported	Start to score this behavior when the rat puts its weight on its hind legs, raises its forepaws from the ground, and extents its head upward. Its forepaws can either lean on the wall or stay suspended
Inactive	The rat will be sitting still on the floor, without performing any of the other scored behaviors, and showing from little to no movement based on the rat's body center-point
Walking	Start behavior when the rat's body center-point begins to move

Data management

The data were archived on the cloud platform of Sylics (Synaptologics; Amsterdam, The Netherlands) which allowed us to share video files, raw data, and spreadsheets between the three sites in an efficient and secure way.

Statistics

The behavioral outcomes were analyzed using SPSS according to the different objectives defined:

a. Reproducibility across sites

A three-way ANOVA with genotype (two levels) and site (three levels) as between-subject factors, and treatment as the repeated within-subject factor [four levels (vehicle, 3 doses)] was performed on absolute data values for each of the readouts. In the case of a main site and genotype effect in this absolute data set, normalized values (relative to vehicle treatment) were analyzed using the same three-way ANOVA design. This analysis aimed to address reproducibility across sites in terms of the phenotype evaluation as well as the pharmacological intervention.

b. Method of scoring

To explore the effect of different methods of scoring, manual scored data was compared to automated data using the same three behavioral outcomes (walking, rearing and circling). Only the middle 10-minute bin of the entire 30-minute observation period was analyzed by employing a 4-way ANOVA with genotype and site as between-subject factors, and method of scoring [two levels: manual (The Observer XT 13) and automatic (EthoVision XT 12)] and treatment as repeated within-subject factors.

RESULTS

Standardization across sites

To ensure high-quality data, the protocol shared across sites addressed randomization and blinding principles in addition to detailed environmental variables, handling and testing procedures summarized in Table 3. The experimental protocol for this study was based on a previous single-site study using the same compound and different automated scoring equipment [14, 17]. The study design was then discussed between the consortium partners to address alignment of factors with anticipated higher relevance to maximize the power of the study. A summary of aligned and non-aligned factors is presented in Table 3.

Table 3. Summary of the experimental factors that were aligned across sites.

Factor	RUG	Pfizer	Roche	Aligned?
Provider of animals	Charles River	Charles River	Charles River	Y
Age at start of experiment	~3 months	~3 months	~3 months	Y
Average bodyweight at the start of experiment	WT 407gr (±33)/ KO 399gr (±45)	300-350gr	WT 428 gr (±32)/ KO 377gr (±33)	N
Animal-related guidelines				
Housing	Single housed	Single housed	Single housed	Y
Cage size	Makrolon type 2L	Innovive Rat Cage	Makrolon type IV	N
Bedding	Lignocel BK8/15	Alpha Dri	Lignocel FS-14	N
Food type	Standard Altronim rodent chow	Standard Purina rat chow (5053)	KLIBA NAFAG 3436	N
Cage cleaning	1/week Friday	1/week Friday	1/week Friday	Y
Enrichment	Wooden bar, nesting material (Enviro-dry)	Plastic bone, nesting material (Bed-R'Nest)	Wooden bar, nesting material	Y[#]
Handling	Tail	tail	Body	N
Experimenter	Master student (MAA)	Undergraduate Researcher (MCD)	PI (BB)	N
Gloves	Yes	Yes	Yes	Y
Disturbance	other rats housed	Not applicable	Radio (60dB)	N
Identification	Cage card, ear-clip and tail mark	Cage card and tail mark	Cage card	N
Physical environment of the housing room (HR)				
Humidity	42%	45%	50%	N
Temperature	73^{0}F	72^{0}F	70^{0}F	N
Lighting	~35 Lux	~35 Lux	~150 Lux	N
Behavioral testing				
Testing days	Mon-Thur.	Mon-Thur.	Mon-Thur.	Y
Test room	Separate	Same as HR	Same as HR	N
Lighting in procedure room	~35 lux	~35 Lux	~150 Lux	N
Volume	35-40 dB	60 dB	60dB	N
Temperature	73 °F	72 °F	70 °F	N
Humidity	42%	45%	50%	N
Randomization	Latin-square	Latin-square	Latin-square	Y
Sample size per genotype	11 WT/12 KO	12 WT/12 KO	12 WT/12 KO	Y*
Dosing	s.c. in the flank	s.c. in the flank	s.c. in the flank	Y
	in holding room	in procedure room	in procedure room	N
Post-dosing time	30 min	30 min	30 min	Y
Environment and Equipment				
PC	outside procedure room	in procedure room	in procedure room	N

Table 3. Summary of the experimental factors that were aligned across sites. (continued)

Factor	RUG	Pfizer	Roche	Aligned?
Experimenter present	No	yes, behind blinders	Yes	N
Equipment	PhenoTyper 4500	PhenoTyper 4500	PhenoTyper 4500	Y
Light inside PhenoTyper	~14.5 Lux	~14 Lux	~80 Lux	N**
Cleaning	alcohol wipes	alcohol wipes	alcohol wipes	Y
Software				
Automated scoring	EthoVision XT 12	EthoVision XT 12	EthoVision XT 12	Y
Manual scoring	The Observer XT 13	The Observer XT 13	The Observer XT 13	Y
Blinded scoring	yes	yes	Yes	Y
Compound				
Provider	Janssen Pharmaceutica NV	Janssen Pharmaceutica NV	Janssen Pharmaceutica NV	Y

*One rat missed the last dosing so it was excluded from the analyses. #home cage enrichment was agreed to be applied at all three sites but adhered to institutional standard practices.
**light intensity inside the chamber was aligned but technical issues prevented one of the sites to use the agreed intensity.

Behavioral evaluation

The behavioral read-outs were 1) circling behavior, expressed as the frequency of circling (clockwise and counterclockwise), 2) rearing, expressed as frequency of supported and unsupported rearing, and 3) time spent walking. These behavioral read-outs were analyzed in separate ANOVA's.

Hyperactive and repetitive phenotypes of *Shank*2 KOs were consistently observed across study sites

Analysis of the automated scorings (EthoVision XT 12), during the 30-minute PhenoTyper chamber exposure, revealed that *Shank*2 KO rats showed increased walking (Figure 1A; *genotype F(1,65)=94.95, p<0.001*), rearing (Figure 1B; *genotype F(1,65)=35.9, p<0.001)*, and circling behavior (Figure 1C; *genotype F(1,65)=22.69, p<0.001*) relative to the WTs across all three study sites. However, a significant genotype x site interaction, and genotype x treatment interaction effect was observed for walking (*F(2,65)=5.9, p<0.005; (F(3,195)=29.9, p<0.001)* and circling (*F(2,65)=3.0, p<0.05; F(3,195)=5.6, p<0.001*) while for rearing only the latter interaction reached significance (F(3,195)=13.5,p<0.001). In addition, a significant overall site effect for rearing (*F(2,65)=7.1, p<0.005*) and circling *(F(2,65)=3.6, p<0.05*) was found. Univariate ANOVA of only the vehicle data showed a significant genotype effect for walking, rearing and circling across all three sites (Figure 1A-C). To display the hyperactive and repetitive behavioral phenotype of *Shank*2 KO rats, the vehicle data from all three sites were pooled for both phenotypes (Figures 1J-L) and analyzed with ANOVA (walking: F(1,70)=48.9, p<0.001; rearing: F(1,69)=31.8, p<0.001; circling: F(1,70)= 55.6, p<0.001).

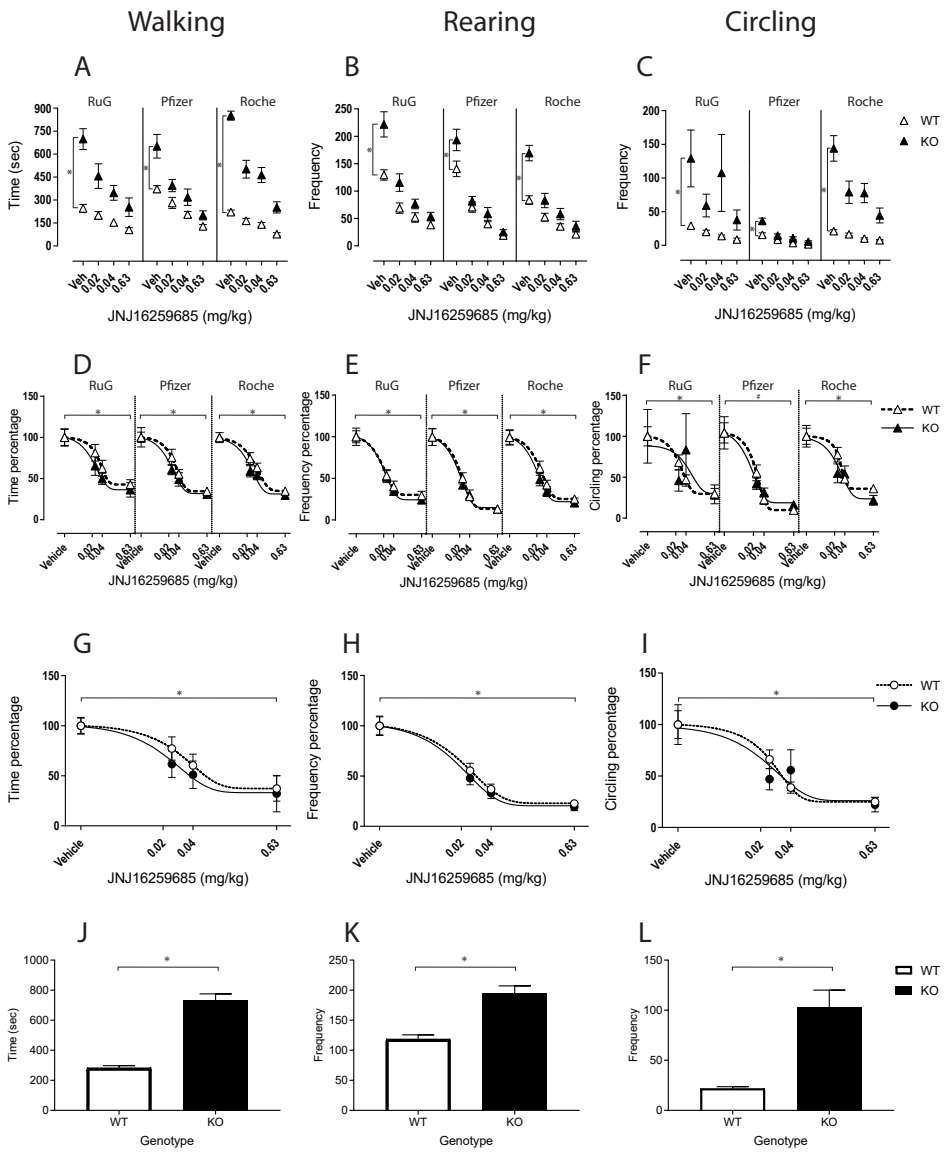


Figure 1: A-C) Absolute values for walking (A), rearing (B), and circling (C) across the three sites from automated scoring analysis (full 30' session). D-F) Normalized values relative to the vehicle for walking (D), rearing (E), and circling (F) across the three sites. The averaged site values for the respective dose levels (depicted in D-F) are shown in G-I. Pooled data from the three sites under vehicle condition for both genotypes for walking (J), rearing (K), and circling (L) behavior. All values are expressed as mean ± S.E.M. *statistically significant differences ($p<0.05$), # points out trend ($p=0.06$).

Further analyses suggested that the site x genotype interaction effect for walking was explained by the higher Pfizer scores of the WT group compared to the RUG and Roche values (*F(2,32)=13.98, p<0.001*). The site effect for rearing was due to the lower Roche scoring values of both groups of animals compared to the RUG and Pfizer data sets *(F(2,65)=7.12, p<0.005*). The site and site x genotype interaction effect for circling was mainly caused by the general lower Pfizer values of, primarily, the *Shank*2 KO group (*F(2,33) =5.19, p<0.02*) and mildly by the WT (*F(2,32) =22.58, p<0.001*) compared to the RUG and Roche values. Yet, all these study-site and genotype main- and interaction-effects disappeared when normalizing the raw data by expressing them as relative to the vehicle control (Figure 1D-F, supplementary Table 2).

Consistent dose-dependent attenuation of motor activity and circling behavior in *Shank2* KO and WT rats by JNJ16259685 treatment across study sites

JNJ16259685 treatment resulted in a robust dose-dependent suppression of walking, rearing, and circling behavior in both WT and *Shank*2 KO rats at all three study sites. For all three behavioral parameters, a significant overall main effect of treatment was found (walking *(Fig 1A)*: *F(3,195)=125.3, p<0.001),* rearing *(Fig 1B): F(3,195)=192.6, p<0.001),* and circling *(Fig 1C): F(3,195)=12.19, p<0.001*) as well as a significant treatment x genotype interaction (Supplementary Table 2). This interaction effect is predominantly caused by the robustly enhanced hyperactivity and repetitive circling behavior of the KO animals and completely disappears when normalizing the raw data by expressing them as relative to the vehicle condition (Figure 1: D-F, Supplementary Table 2). This indicates that the JNJ16259685 treatment effects were similar for KO and WT rats and, importantly, consistent across all three sites. Combining the data from all three sites (Fig 1G-I) demonstrated similar dose-response curves for JNJ16259685 treatment to inhibit walking (ID_{50} = 0.9549 and 0.583 for WTs and KOs, respectively), rearing (ID_{50} = 0.6755 and 0.6883 for WTs and KOs respectively), and circling behavior (ID_{50} = 1.034 and 0.535 for WTs and KOs respectively) for both WT and *Shank*2 KO animals.

Treatment effects are comparable whether scored automatically or manually

The present study was focusing on the reproducibility of automated scored behavior from a previous study [17]. To test whether the level of reproducibility of the automated scoring was comparable to that of manual scoring, a method frequently used in behavioral pharmacology studies, we compared manual and automated scored behaviors in a 10-minute segment of the data across sites. Manual (The Observer XT 13) and automated (EthoVision XT 12) scorings of the second 10-minute bin of the recordings were employed and included in the ANOVA as an additional (within-subject) factor;

this 10-minute time segment was selected because it had the highest rate of activity in the 60-minute evaluation of Modi et al., 2018. A four-way ANOVA analysis revealed a significant main effect of method for all three behavioral parameters (see Figures 2A-C for walking; Figures 2D-F for rearing, and Figures 2G-I for circling, and Supplementary Table 3) as well as several method interaction effects with genotype, site, and treatment (see Supplementary Table 3).

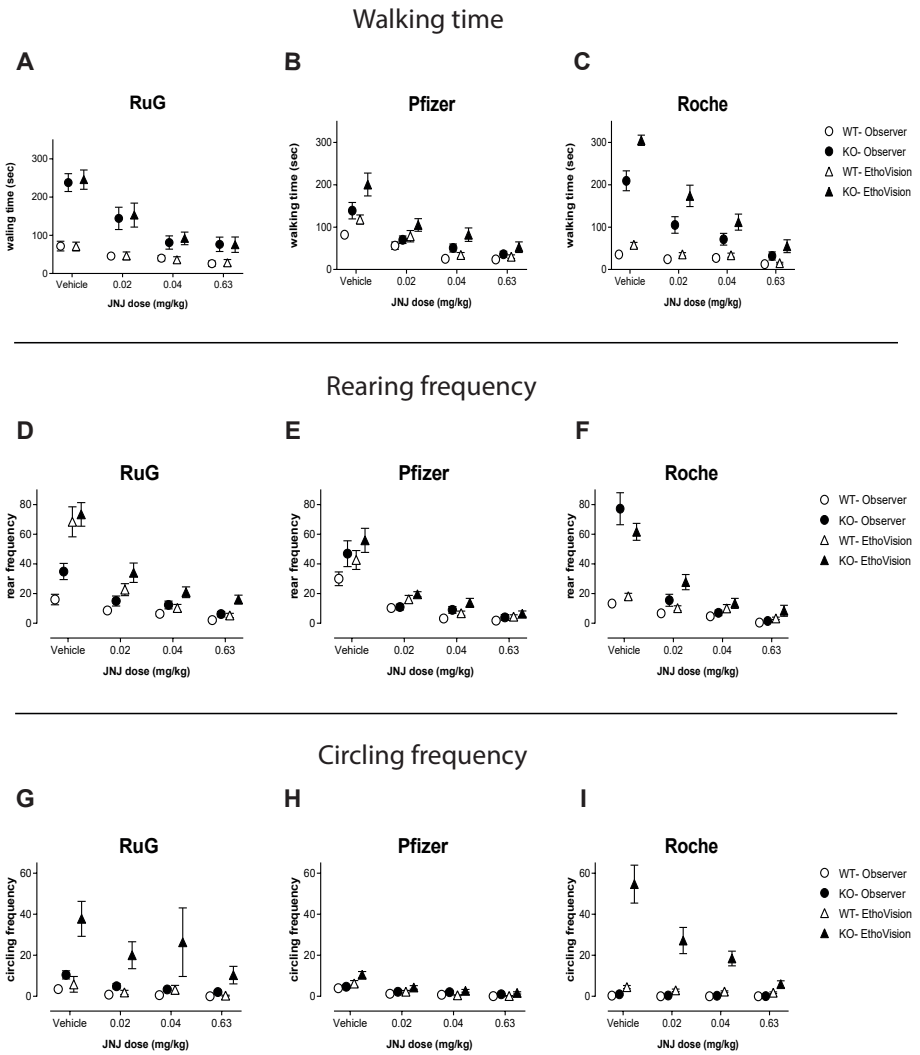


Figure 2. Absolute values across treatments for walking time (A-C), rearing frequency (D-F), and circling frequency (G-I) for the second 10-minute bin of the automated (triangles) and manual scored (circles) data, comparing the KO (black shapes) and WT (white shapes) rats. All data points express the mean ± S.E.M.

Overall, the automated EthoVision-scored values for walking, rearing, and circling were higher than the manual Observer-scored values (*F(1,65)=112.1,* p<0.001; *F(1,65)=47.8, p<0.001; and F(1,65)=52.9, p<0.001*; respectively). The hyperactive and repetitive circling behavioral phenotype of the KOs as well as its dose-dependent suppression by JNJ16259685 treatment were reliably detected at all three study sites by manual scoring (Walking: Fig 2A-C; Rearing: Fig 2D-F; Circling: Fig 2G-I). Interestingly, while manual- and automated-scored circling displayed visibly different scores across sites, no significant main or interaction effects appeared following normalization of the raw data relative to the vehicle condition. However, for the normalized rearing data for the 10-minute bin (data not shown), a significant main effect of method and method x treatment interaction persisted (Supplementary Table 3). Further analyses revealed higher values at the JNJ16259685 0.04 and 0.63 mg/kg treatments for the automated method relative to the manual scoring method.

DISCUSSION

The present study demonstrates that rigorous alignment of experimental protocols between three research centers resulted in comparable experimental findings across sites for both genotype and treatment effects. The phenotypic difference between the *Shank2* KO and WT rats was reliably observed in all three study sites. While there were differences in amplitude of the genotype effect on behavior, all three sites observed that the KO rats displayed consistently heightened motor activity (i.e. walking and rearing) and stereotypic circling behavior when compared to WT rats. This finding indicates reproducible findings across sites, in addition to the replication of the original report [17]. Importantly, our data demonstrates the robustness of the *Shank2* deletion-induced behavioral phenotypes that may mimic some of the behavioral abnormalities observed in ASD. Likewise, a consistent and dose-dependent attenuation of motor activity and circling behavior in both KO and WT rats by JNJ16259685 was found across the three study sites.

This high level of reproducibility is likely to be attributed to the rigorously standardized experimental protocol. The study design employed herein was adapted from the original work of Modi et al., 2018 with particular effort to prevent bias in the design, collection, and analysis of data (e.g. blinding, randomization, carry-over effects, etc.); and to analyze the data similarly through automated scoring. Besides the study design, experimental conditions that may have biological relevance to the expression of the phenotype (e.g., age of the animals) were aligned across sites. Conversely, factors that were not expected to have direct biological relevance related to phenotype expression

were addressed; however, at a variable level between sites (e.g., environmental enrichment applied for all sites, but the level of environmental enrichment for the housing conditions differed across sites). Thus, environmental variability between the three sites was allowed, which introduced heterogenization for experimental conditions that were site specific [18, 27]. Therefore, our study appears to support the assumption that a combination of standardized and heterogenized factors can lead to a high level of reproducibility between different laboratories. Selection of these factors may be dependent on study aim and neurobiological construct that is being investigated; indeed, for future study designs, it is recommended to carefully review the standardization of environmental factors and consider their relevance in light of the phenotype of interest. For example, by over-standardizing only factors that are not biologically relevant to the expression of the (behavioral) phenotype of interest, the result is at risk of being highly idiosyncratic. On the other hand, and as recently suggested, introducing systematic heterogenization of certain factors can boost external validity and thus reproducibility [30].

Preclinical studies are a stepping stone in the pipeline for new pharmacotherapeutic treatments of human disorders. Thus, the development and assessment of animal models that recapitulate specific phenotypes of the disorders in a consistent manner is crucial when testing new therapeutic targets. In addition to protocol alignment for factors related to the laboratory (micro-)environment, selection of the type of animal model is also important in view of reproducibility (e.g., when the originally observed effect sizes in outcome measures for the selected model are small). Here, the initial hyperactive and repetitive behavioral phenotypes of *Shank2* KO rats were robust as these behavioral alterations are consistently observed across various different *Shank*-mutations in both rats and mice under a variety of experimental testing conditions [6, 12, 16, 19, 22, 25] suggesting the authentic relevance of these postsynaptic scaffolding proteins that are present at glutaminergic synapses for ASD-like behaviors and the suitability for pharmacological testing. Nonetheless, attention must be drawn to the different underlying circuitry responsible for the robust phenotype since it might not completely overlap across the different *Shank* mutations, as previously reported by Yoo et al. 2014 who found inconsistencies in molecular, physiological and behavioral data between and within the *Shank* mutant mouse lines[31].

Recapitulating and expanding on the findings from Modi et al. (2018), administration of the selective and high-affinity mGluR1 antagonist JNJ16259685 effectively attenuated the hyperactivity and repetitive circling behavior of *Shank*2 KO rats in a dose-dependent manner. While Modi et al. demonstrated a significant attenuation of these behavioral phenotypes in both WT and KO animals, they argued that JNJ16249685 (0.63 mg/kg) normalized KO behavior to WT vehicle-dosed levels. Here, we show that the

locomotor-suppressing effects of JNJ16259685 produce similar dose-effect curves in both genotypes. This goes well in line with the fact that the mGluR1 receptors are richly distributed in regions associated with motor function including the cerebellum [7, 20] and basal ganglia [4], and are believed to play an important role in movement, motor coordination, and motivation [1, 11]. Our findings agree with the results of Hodgson et al., 2011, who reported a dose-dependent reduction in novelty-induced locomotor and rearing activity of Wistar rats. Hence, they support the assertion that the mGluR1 is involved in general motivation to explore their environment [11]. Although this study was focusing on reproducing the behavioural features in the Shank2 KO rats, electrophysiological characterization can be reviewed in Modi, et al. (2018). Overall, our results support the suggestion that the hyperactive phenotype of *Shank2*-deficient rats is associated with enhanced striatal mGluR1 signaling [17] thereby providing face, construct and predictive validity of this animal model for ASD.

Another aim of this study was to compare two different methods of scoring behavior, automated versus manual scoring. Behavioral studies are relevant for most biological, evolutionary, and biomedical research questions, creating a need for high-throughput experiments and mechanistic insight; however, the effort and time spent in manual scoring and data processing becomes a burden when conducting behavioral experiments. Therefore, there is a need for an automated screening of an animal's behavior capable to discriminate between different behavioral categories, especially in the presence of animal manipulations. For the current study, three behavioral categories were chosen to compare between an automated and a manual scoring; these categories have a different level of complexity in terms of how straightforward it is to score the behavior. The selected categories are walking time, rearing frequency and circling frequency, from the most to least simple.

Overall, the automated scoring showed higher rates compared to the manual scoring at all three sites. The mismatch between methods was present for both genotypes and across treatments indicating that the differences between methods might originate from the flexibility of the behavioral definition adopted for each scoring method; in addition, discrepancies might also be attributed to the human observer 'smoothing' the scoring, meaning that brief intervals between behaviors are scored as the continuity of the behavior, while the automated scoring counts separate events. This suggests that special attention must be drawn not only to the definition of the behavior being scored, but also to the parameters that frame this definition, likely in this case smoothing by the human scorer and the continuity of the behavior scored as separate events by the automated scoring. These parameters have to be adapted according to the behavior being defined and the instrument used. Moreover, the concordance between methods was

higher for the simplest behavioral category (walking time) and the lowest for the most complex category (circling frequency) suggesting that the coherence between scoring methods is more easily attainable when the behavioral category is unambiguous. Importantly, both the manual and automated scoring methods succeeded in detecting the phenotype and treatment effects (Supplementary Table 3), suggesting that they are both reliable methods to assess the relatively simple behaviors scored in the current study.

To conclude, by using a combination of standardization and heterogenization for experimental factors, a harmonized protocol was generated and applied to a multicenter study in which genotype and treatment effects were studied at a behavioral level. Here we showed that, following careful alignment of these factors, reproducibility of genotype and treatment effects in rodents can be established for both automated- and manually-scored behaviors. The present study also reveals important implications for the treatment efficacy of mGluR1 antagonism for a translational phenotype related to a major risk gene for Autism Spectrum Disorders.

ACKNOWLEDGEMENTS

The support of Heidi Huysmans (Janssen Pharmaceutica NV) in the logistics of compound shipping is highly appreciated.

The present study was supported by the European Autism Interventions - A Multicenter Study for Developing New Medications (EU-AIMS) project, which receives support from the Innovative Medicines Initiative Joint Undertaking under Grant agreement number 115300, resources of which are composed of financial contributions from the European Union's Seventh Framework Programme (FP7/2007-2013), from the European Federation of Pharmaceutical Industries and Associations companies' in-kind contributions, and from Autism Speaks.

COMPETING INTEREST STATEMENT

The authors declare no competing interests. During the study, R.G, M.C-D, and D.L.B. were full time employees and shareholders of Pfizer, Inc. D.L.B. is currently a fully time employee and shareholder of Takeda Pharmaceutical Company, Ltd. M.M., W.S. and B.B are fully employed by Roche. E.v.D., W.v.D., and L.P.J.J.N are fully employed by Noldus Information Technology. E.S. is full time employee of Servier. W.D. is fully employed by

Janssen Pharmaceutica NV and holds stocks and options. B.K and M.L are fully employed by Sylics. LN is the majority shareholder of Noldus Information Technology BV.

AUTHOR CONTRIBUTION STATEMENT

All authors were involved in the study design. They have all reviewed and approved the manuscript before submission. M.A-A., R.G., and M.M. conducted the experiments and data analysis. M.C-D conducted the experiments. B.B., D.L.B., and M.J.K. performed the overall supervision of the project for each study site. M.A-A., S.F.d.B., and M.J.K. wrote the initial version of the manuscript.

REFERENCES

1. Aiba, A., Kano, M., Chen, C., Stanton, M. E., Fox, G. D., Herrup, K., Tonegawa, S. (1994). Deficient cerebellar long-term depression and impaired motor learning in mGluR1 mutant mice. Cell, 79(2), 377–388.
2. Baker, J. D. (2016). The Purpose, Process, and Methods of Writing a Literature Review. AORN Journal, 103(3), 265–269. https://doi.org/10.1016/j.aorn.2016.01.016
3. Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. Nature News, 505(7485), 612. https://doi.org/10.1038/505612a
4. Conn, P. J., Battaglia, G., Marino, M. J., & Nicoletti, F. (2005). Metabotropic glutamate receptors in the basal ganglia motor circuit. Nature Reviews Neuroscience, 6(10), 787–798. https://doi.org/10.1038/nrn1763
5. Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: interactions with laboratory environment. Science (New York, N.Y.), 284(5420), 1670–1672.
6. Ergaz, Z., Weinstein-Fudim, L., & Ornoy, A. (2016). Genetic and non-genetic animal models for autism spectrum disorders (ASD). Reproductive Toxicology (Elmsford, N.Y.), 64, 116–140. https://doi.org/10.1016/j.reprotox.2016.04.024
7. Fotuhi, M., Sharp, A. H., Glatt, C. E., Hwang, P. M., von Krosigk, M., Snyder, S. H., & Dawson, T. M. (1993). Differential localization of phosphoinositide-linked metabotropic glutamate receptor (mGluR1) and the inositol 1,4,5-trisphosphate receptor in rat brain. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 13(5), 2001–2012.
8. Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The Economics of Reproducibility in Preclinical Research. PLOS Biology, 13(6), e1002165. https://doi.org/10.1371/journal.pbio.1002165
9. Gerlai, R. (2018). Reproducibility and replicability in zebrafish behavioral neuroscience research. Pharmacology, Biochemistry, and Behavior. https://doi.org/10.1016/j.pbb.2018.02.005
10. Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? Science Translational Medicine, 8(341), 341ps12-341ps12. https://doi.org/10.1126/scitranslmed.aaf5027
11. Hodgson, R. A., Hyde, L. A., Guthrie, D. H., Cohen-Williams, M. E., Leach, P. T., Kazdoba, T. M., … Varty, G. B. (2011). Characterization of the selective mGluR1 antagonist, JNJ16259685, in rodent models of movement and coordination. Pharmacology, Biochemistry, and Behavior, 98(2), 181–187. https://doi.org/10.1016/j.pbb.2010.11.018
12. Jiang, Y., & Ehlers, M. D. (2013). Modeling Autism by SHANK Gene Mutations in Mice. Neuron, 78(1), 8–27. https://doi.org/10.1016/j.neuron.2013.03.016
13. Kafkafi, N., Agassi, J., Chesler, E. J., Crabbe, J. C., Crusio, W. E., Eilam, D., … Benjamini, Y. (2018). Reproducibility and replicability of rodent phenotyping in preclinical studies. Neuroscience and Biobehavioral Reviews, 87, 218–232. https://doi.org/10.1016/j.neubiorev.2018.01.003
14. Lavreysen, H., Wouters, R., Bischoff, F., Nóbrega Pereira, S., Langlois, X., Blokland, S., … Lesage, A. S. J. (2004). JNJ16259685, a highly potent, selective and systemically active mGlu1 receptor antagonist. Neuropharmacology, 47(7), 961–972. https://doi.org/10.1016/j.neuropharm.2004.08.007
15. Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. Science, 355(6325), 584–585. https://doi.org/10.1126/science.aal3618
16. Mei, Y., Monteiro, P., Zhou, Y., Kim, J.-A., Gao, X., Fu, Z., & Feng, G. (2016). Adult restoration of Shank3 expression rescues selective autistic-like phenotypes. Nature, 530(7591), 481–484. https://doi.org/10.1038/nature16971

17. Modi, M. E., Brooks, J. M., Guilmette, E. R., Beyna, M., Graf, R., Reim, D., … Buhl, D. L. (2018). Hyperactivity and Hypermotivation Associated With Increased Striatal mGluR1 Signaling in a Shank2 Rat Model of Autism. Frontiers in Molecular Neuroscience, 11. https://doi.org/10.3389/fnmol.2018.00107
18. Richter, S. H., Garner, J. P., Auer, C., Kunert, J., & Würbel, H. (2010). Systematic variation improves reproducibility of animal experiments. Nature Methods, 7(3), 167–168. https://doi.org/10.1038/nmeth0310-167
19. Schmeisser, M. J., Ey, E., Wegener, S., Bockmann, J., Stempel, A. V., Kuebler, A., … Boeckers, T. M. (2012). Autistic-like behaviours and hyperactivity in mice lacking ProSAP1/Shank2. Nature, 486(7402), 256–260. https://doi.org/10.1038/nature11015
20. Shigemoto, R., Nakanishi, S., & Mizuno, N. (1992). Distribution of the mRNA for a metabotropic glutamate receptor (mGluR1) in the central nervous system: an in situ hybridization study in adult and developing rat. The Journal of Comparative Neurology, 322(1), 121–135. https://doi.org/10.1002/cne.903220110
21. Steckler, T. (2015). Editorial: preclinical data reproducibility for R&D - the challenge for neuroscience. SpringerPlus, 4(1), 1. https://doi.org/10.1186/2193-1801-4-1
22. Vicidomini, C., Ponzoni, L., Lim, D., Schmeisser, M. J., Reim, D., Morello, N., … Verpelli, C. (2017). Pharmacological enhancement of mGlu5 receptors rescues behavioral deficits in SHANK3 knockout mice. Molecular Psychiatry, 22(5), 689–702. https://doi.org/10.1038/mp.2016.30
23. Voelkl, B., Vogt, L., Sena, E. S., & Würbel, H. (2018). Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLOS Biology, 16(2), e2003693. https://doi.org/10.1371/journal.pbio.2003693
24. Voelkl, B., & Würbel, H. (2016). Reproducibility Crisis: Are We Ignoring Reaction Norms? Trends in Pharmacological Sciences, 37(7), 509–510. https://doi.org/10.1016/j.tips.2016.05.003
25. Won, H., Lee, H.-R., Gee, H. Y., Mah, W., Kim, J.-I., Lee, J., … Kim, E. (2012). Autistic-like social behaviour in Shank2-mutant mice improved by restoring NMDA receptor function. Nature, 486(7402), 261–265. https://doi.org/10.1038/nature11208
26. Würbel, H. (2000). Behaviour and the standardization fallacy. Nature Genetics, 26(3), 263. https://doi.org/10.1038/81541
27. Würbel, Hanno. (2007). Refinement of rodent research through environmental enrichment and systematic randomization, 9.
28. Würbel, Hanno. (2017). More than 3Rs: the importance of scientific validity for harm-benefit analysis of animal research. Lab Animal, 46, 164–166. https://doi.org/10.1038/laban.1220
29. Beynen A.C., Gartner, K., van Zutphen L.F.M. Standardization of animal experimentation. In: Zutphen LFM, Baumans, V, Beynen AC, editors. Principles of laboratory animal science. 2nd ed. Amsterdam: Elsevier Ltd; 2003. pp. 103-110
30. Bodden, C., Kortzfleisch, V. T. von, Karwinkel, F., Kaiser, S., Sachser, N., & Richter, S. H. (2019). Heterogenising study samples across testing time improves reproducibility of behavioural data. Scientific Reports, 9(1), 8247.
31. Yoo, J., Bakes, J., Bradley, C., Collingridge, G. L., & Kaang, B.-K. (2014). Shank mutant mice as an animal model of autism. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1633). https://doi.org/10.1098/rstb.2013.0143

Systematic assessment of the replicability and generalizability of preclinical findings: impact of protocol harmonization across laboratory sites.

María Arroyo-Araujo[1], Bernhard Voelkl[2], Clément Laloux[3], Janja Novak[2], Bastijn Koopmans[4], Ann-Marie Waldron[6], Isabel Seiffert[6], Helen Stirling[6], Katharina Aulehner[6], Sanna K. Janhunen[7], #a, Sylvie Ramboz[5], Heidrun Potschka[6], Johanna Holappa[7], Tania Fine[8], Maarten Loos[4], Bruno Boulanger[3], Hanno Würbel[2], and Martien J. Kas[1,*]

[1]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands
[2]Animal Welfare Division, Vetsuisse Faculty, University of Bern, Bern, Switzerland
[3]Pharmalex, Mont-Saint-Guibert, Belgium
[4]Sylics (Synaptologics BV), Amsterdam, The Netherlands
[5]PsychoGenics Inc., New Jersey, Paramus, United States of America
[6]Inst. of Pharmacology, Toxicology, and Pharmacy, Ludwig-Maximilians-Universitaet Muenchen, Muenchen, Germany
[7]Orion Pharma, Turku, Finland
[#]a Current address: Organon R&D Finland, Turku, Finland
[8]Teva Pharmaceuticals, Tel Aviv, Israel
[*]m.j.h.kas@rug.nl

Published in PLoS Biol. 2022 Nov 23;20(11):e3001886.

ABSTRACT

The influence of protocol standardization between laboratories on their replicability of preclinical results has not been addressed in a systematic way. While standardization is considered good research practice as a means to control for undesired external noise (*i.e.*, highly variable results) some reports suggest that standardized protocols may lead to idiosyncratic results thus undermining replicability. Through the EQIPD consortium, a multi-lab collaboration between academic and industry partners, we aimed to elucidate parameters that impact the replicability of preclinical animal studies. To this end, three experimental protocols were implemented across 7 laboratories. The replicability of results was determined using the distance travelled in an open field after administration of pharmacological compounds known to modulate locomotor activity (MK-801, diazepam, and clozapine) in C57BL/6 mice as a worked example. The goal was to determine whether harmonization of study protocols across laboratories improves the replicability of the results and whether replicability can be further improved by systematic variation (heterogenization) of two environmental factors (time of testing and light intensity during testing) within laboratories. Protocols were tested in three consecutive stages and differed in the extent of harmonization across laboratories and standardization within laboratories: *stage 1*- minimally aligned across sites (local protocol), *stage 2*- fully-aligned across sites (harmonized protocol) with and without systematic variation (standardized and heterogenized cohort), and *stage 3*- fully-aligned across-sites (standardized protocol) with a different compound. All protocols resulted in consistent treatment effects across laboratories, which were also replicated within laboratories across the different stages. Harmonization of protocols across laboratories reduced between-lab variability substantially compared to each lab using their local protocol. In contrast, the environmental factors chosen to introduce systematic variation within laboratories did not affect the behavioral outcome. Therefore, heterogenization did not reduce between-lab variability further compared to the harmonization of the standardized protocol. Altogether, these findings demonstrate that subtle variations between lab-specific study protocols may introduce variation across independent replicate studies even after protocol harmonization and that systematic heterogenization of environmental factors may not be sufficient to account for such between-lab variation. Differences in replicability of results within and between laboratories highlight the ubiquity of study-specific variation due to between-lab variability, the importance of transparent and fine-grained reporting of methodologies and research protocols, and the importance of independent study replication.

INTRODUCTION

In recent years, the scientific community has raised concerns about the replicability of results, particularly in the preclinical biomedical sciences. Defining results replicability as the ability to duplicate results from a previous scientific claim supported by new data (1,2). Various causes of poor replicability have been proposed, including the diverse methodologies used in the field and the lack of rigorous research practices (*e.g.,* underpowered studies, risks of biases, inadequate statistics) (3–7). Although these causes can certainly explain part of the problem, they permeate different science subfields differently (8) and cannot account for the poor replicability of results on their own. To our knowledge, no systematic studies have been performed to investigate the effect of protocol standardization within laboratories and protocol harmonization across laboratories regarding between-laboratory variation in light of replicability and generalizability of results.

The current and most common research practice of conducting single laboratory studies under standardized conditions has recently been proposed as a source of the high variability of results between laboratories (9,10). Whenever rigorous standardization of environmental conditions within a study leads to homogenous study populations, the study results may become idiosyncratic as the study population is only representative of the narrow set of conditions in which it was tested. This increases the risk of replication failure even under only slightly different conditions as standardized; such single-site study designs do not allow predicting changes in the expression of the phenotype in response to different environmental influences. The change in the expression of the phenotype is caused by biological variation (11) which describes how genetic variation interacts with environmental factors to which experimental animals are exposed throughout development (gene-environment interactions), thereby shaping their phenotype (12).

Another approach taken to deal with the variability of results across laboratories is to harmonize the same standardized protocol across studies (13). If harmonization includes those environmental and experimental factors that may influence the phenotype expression, it should result in replicable results. However, current evidence is ambiguous. Whereas in one study a rigorously standardized protocol that was harmonized across 3 laboratories resulted in many non-replicable findings (14), another study that also followed protocol standardization and harmonization across 3 sites found similar phenotypic and pharmacological effects; however, the proportion of variation explained by lab was not formally assessed (15). This suggests that this experimental approach may be missing to address some unknown source of variability between sites.

Certainly, there are inherent differences between laboratory environments which are not addressed in multi-laboratory protocols because of the low feasibility of harmonizing them or simply because these differences are not known (*e.g.,* different ways to handle the animals, diversity in equipment). Some of these differences likely interact with the phenotype expression; this interaction may be accentuated when other sources of variability are minimized (*i.e.,* standardized). Thus, although the same standardized protocol is implemented in different sites, it may still produce different results (16). Still, there are no accounts to evaluate the impact that protocol harmonization across sites has on between-lab variability.

Furthermore, it has been recently suggested that if the between-lab variation can be incorporated within a single lab, the replicability of results between studies would increase (17–19). Such an approach has been previously implemented (17,19–21) yet, it has not been compared to a non-harmonized study across laboratories to assess the effect on between-lab variation.

To shed light on the effects of protocol harmonization across laboratories, we studied on one side whether harmonization of a standardized protocol reduces between-lab variation in comparison to a non-harmonized local protocol. Furthermore, we tested the effect of systematic heterogenization to assess whether within-lab heterogenization can further reduce between-lab variation compared to the standardized protocol. The experiments performed in this paper are defined as knowledge-claiming research according to Bespalov, et al., (2021) (22).

RESULTS

Stage 1: Local protocol

In this stage, two different compounds with opposite effects were tested to assess their effect on the distance traveled in the OF across the seven sites. The 3 mg/kg Diazepam group showed strong sedative effects (*i.e.,* no distance traveled) relative to its control group; this made the comparison across treatments and laboratories uninformative given the floor effect (Tables A and B in S1 Supplementary Stage). Therefore, the analysis of results was focused on the effects of MK-801.

The local protocol showed a significant drug treatment effect with 0.2 mg/kg MK-801 increasing locomotion compared to saline treatment; this was replicated across all sites (Fig 1). When looking closely at this effect, although all sites found a significant effect, effect size differed across sites. On the other hand, the treatment with 0.3 mg/kg MK-

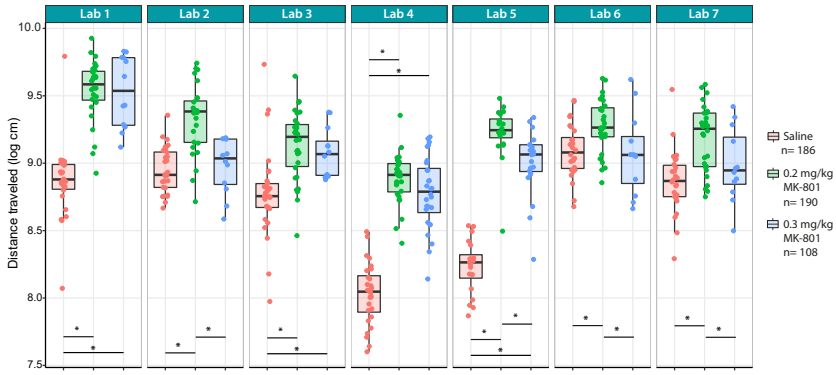


Fig 1. Tukey box plots and individual data points of the total distance traveled (log-transformed) in a 15-minute open field across 7 laboratories. All labs reported significant differences (*p< 0.05) between the groups receiving saline (red symbols) and 0.2 mg/kg MK-801 (green symbols). Labs 1, 3, 4 and 5 also found significant differences between animals receiving 0.3 mg/kg MK-801 (blue symbols) and saline. In addition, labs 2, 5, 6 and 7 found significant differences between the two different drug treatments of MK-801. Data underlying this figure can be found in https://osf.io/8f6yr/, Stage 1 folder.

801 drug treatment on distance moved for all seven sites were into the same direction; however, based on statistical findings, only 4 out of 7 sites found a significant increase in distance moved (Fig 2). All the statistical results of the analysis of the treatment effect per laboratory can be found in Table A in S1 Supplementary Stage.

The comparison of results across laboratories (model 2) revealed that one-third (33%) of the total variance was associated with differences between laboratories. The interaction of the drug treatment effects and the laboratory explained 25% of the variance while the remaining 41% of the variance was attributed to the residual (Table 1).

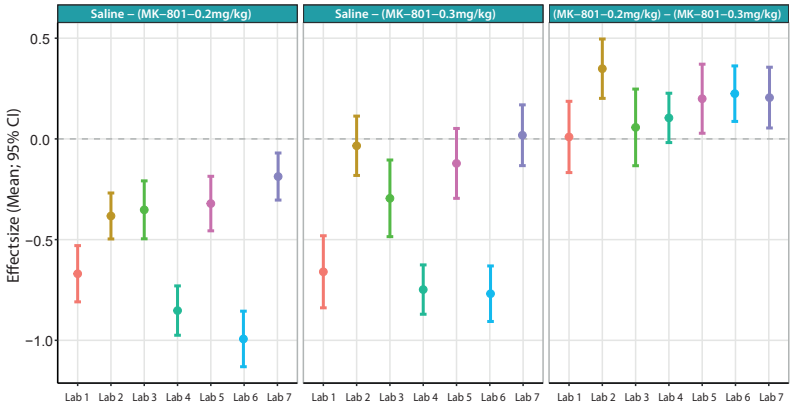


Fig 2. Mean and 95% CI of the treatment effect differences across laboratories between saline and MK-801 0.2 mg/kg (left panel), betweenMK-801 0.3 mg/kg (middle panel) and comparing both MK-801 drug treatments with each other (right panel). Data underlying this figure can be found in the Table A in S1 Supplementary Stage.

Stage 2: Local and harmonized (standardized and heterogenized) protocols

This stage aimed to assess the impact of harmonization of protocols across sites, and heterogenization of protocols within sites, on the replicability of the results. Therefore,

Table 1. Variance components of the across-laboratory analysis (model 2).

Parameter	Variance estimate	%
Lab	0.045	33.19
DrugTreatment:Lab	0.034	25.23
Residual	0.057	41.58

standardized and heterogenized cohorts of the harmonized protocol were compared with the local protocol across the 7 sites in terms of between-laboratory variation in results (Fig 3). First, we evaluated the drug treatment effect in each laboratory for each of the protocols (model 1, Tables A, B, and C in S2 Supplementary Stage). When comparing the treatment effects in each of the laboratories, the analysis revealed that for the three protocols the 0.2mg/kg MK-801 resulted in a significantly larger distance traveled than the saline condition.

For each protocol, we found that the variance results from the Local protocol of stage 1 were highly similar to the variance results in stage 2, suggesting replicability when each laboratory followed its own protocol. Variance components for this protocol in both

Table 2. Across-stage comparison of the Local protocol in stages 1 and 2 (model 3).

Parameter	Std. Dev.	Variance Estimate	%
DrugTreatment:Stage	0.041	0.002	1.12
Stage	0.000	0.000	0.00
DrugTreatment:Lab	0.202	0.041	27.45
Lab	0.228	0.052	34.88
Residual	0.234	0.055	36.55
Total	0.705	0.149	100.00

stages are similar and represent the same proportions (Table D in S2 Supplementary Stage). In addition, the across stages model (model 3) shows that the variability induced by the stage is nearly 0 (Table 2).

Across-lab harmonization (standardized cohort) reduced the overall data variability (i.e., Total variance) compared to the Local protocol as summarized in Table 3: "Total" (model 2). Looking at the proportion of variability explained in each protocol we found

that the variance explained by the variability between laboratories ("Lab" in Table 3) in the Harmonized protocol- standardized cohort (18.67%) decreased by a factor of 3.37 compared to the Local protocol. In addition, this cohort also suggests a more replicable treatment effect across participating labs than the Local protocol as it reduced the variance induced by the drug-by-lab interaction ("DrugTreatment:Lab" term) from ~30 to ~7% [29.31% to 7.57%].

The implementation of the Heterogenized cohort of the Harmonized protocol also reduced the overall variance compared to the Local protocol by a factor of 1.8 but was

Table 3. Variance components of the across-lab analysis for the Stage 2 Local, Standardized and Heterogenized protocols (model 2).

Parameter	Local	Standardized	Heterogenized
DrugTreatment:Lab	0.042 (29.31%)	0.006 (7.57%)	0.011 (14.38%)
Lab	0.054 (37.67%)	0.016 (18.67%)	0.024 (30.93%)
Residual	0.048 (33.02%)	0.063 (73.76%)	0.042 (54.69%)
Total	0.144 (100.00%)	0.086 (100.00%)	0.076 (100.00%)

relatively close to the variance from the Standardized cohort (factor 1.1), respectively ("Total" row in Table 3). Taking a closer look at the proportions of explained variance within protocols, we found that the variance associated with the variability across laboratories in the Heterogenized cohort (31%) was slightly reduced compared to the Local protocol (38%) but not actually larger than the Standardized cohort (19%). Similarly, the interaction of the treatment effect by laboratory in the Heterogenized cohort was reduced compared to the Local protocol by a factor of 3.8, though the Standardized cohort led to an even larger reduction (factor of 8.1).

Furthermore, an extra analysis was performed to explore the individual contribution for each of the two environmental factors varied systematically as part of the Heterogenized cohort. The analysis on the light intensity factor revealed that the drug treatment effect across laboratories is not influenced by the light intensity (Tables E and F in S2 Supplementary Stage). Additionally, this factor had no effect on the variability of the measures when included as a random factor in the linear model (Table G in S2 Supplementary Stage).

Similarly, the analysis for the time of testing factor revealed no difference across laboratories for the early vs late time of testing and this factor did not influence the drug treatment effect (Tables H and I in S2 Supplementary Stage) and also had no influence on the variability of the measure (Table J in S2 Supplementary Stage).

Stage 3: Local and harmonized (standardized) protocols

In this stage, we performed three different analyses, each with a different purpose. First,

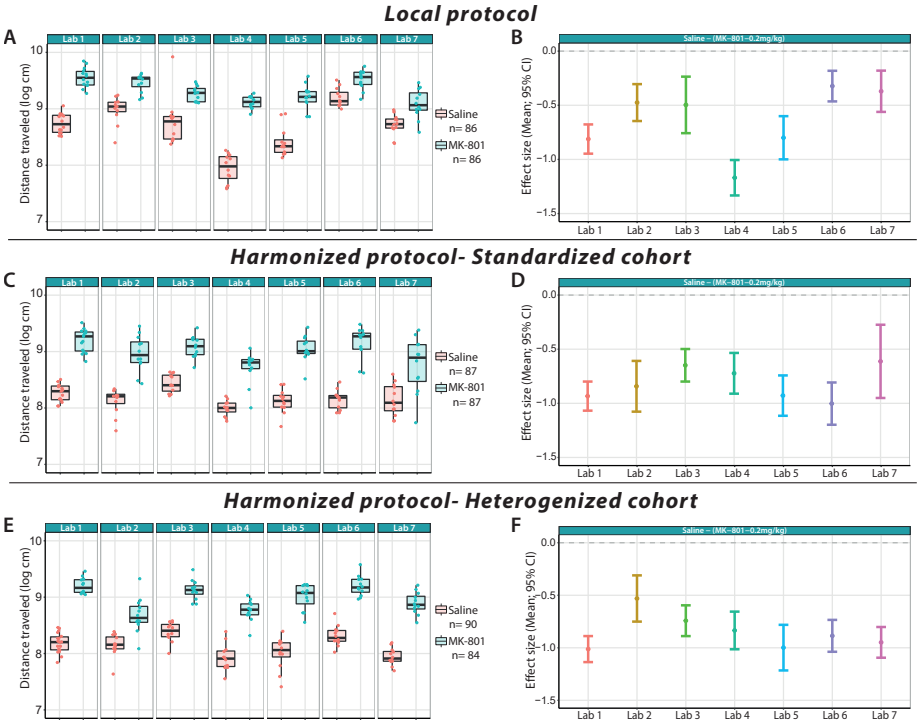


Fig 3. Box-plots and individual data points of the total distance traveled (left column) and treatment effect differences (right column) for the different protocols used in stage 2: A-B: Local, C-D: Harmonized- Standardized cohort, and E-F: Harmonized- Heterogenized cohort. All sites found statistical differences ($p < 0.05$) in the distance traveled after saline (teal symbols) and 0.2 mg/kg MK-801 (red symbols) treatments. Data underlying this panels A, C and E can be found in https://osf.io/8f6yr/, Stage 2 folder. Data underlying panels B, D and F can be found in Supplementary table A, B & C, respectively within S2 Supplementary Stage.

the treatment effect was assessed by comparing the outcome after administration of Clozapine 1 and 2.5 mg/kg, and ultrapure water (Fig 4). Given that this stage was carried out around the contingency of the COVID-19 pandemic, some animal facilities had to stay closed; therefore, Lab 2 was not able to provide data for this stage.

The analysis within each laboratory (model 1) revealed that the Clozapine treatment with 1 mg/kg significantly reduced the distance traveled compared to the ultrapure water for all labs except Lab 1 & 4. However, analysis of the data from these labs revealed a trend with the same direction of the effect (Fig 4: right graph). The highest dose tested,

i.e., 2.5mg/kg Clozapine dose significantly reduced the distance travelled relative to ultrapure water in all sites (Table A in S3 Supplementary Stage).

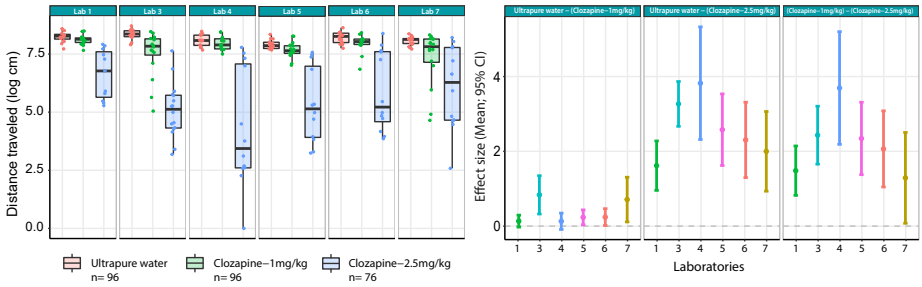


Fig 4. Left graph: Tukey box plots and individual data points across laboratories of the total distance traveled after Clozapine administration (green: 1 mg/kg; blue: 2.5 mg/kg) compared to ultrapure water (red) following the Standardized protocol in stage 3. Right graph: Mean and 95% CI of the treatment effect differences for the Standardized protocol when comparing the control condition to the low dose (left panel), high dose (middle panel) and both doses (right panel) of Clozapine. Data underlying the left panel can be found in https://osf.io/8f6yr/, Stage 3; while data underlying the right panel can be found in Table A in S3 Supplementary Stage.

Secondly, to evaluate the impact of the protocol followed, the between lab variability was compared when following the standardized protocol with the local protocol, both after 2.5 mg/kg Clozapine (model 1). Both, the Local and Standardized protocols showed a similar effect in the distance traveled after Clozapine treatment, although it differed across sites (Fig 5). However, the Standardized protocol reduced the overall variance compared to the Local protocol for this particular treatment (Table 4). In addition, the proportion of the variance explained by the variability across labs when implementing this protocol was reduced by a factor of 2.6 for the 2.5 mg/kg dose of Clozapine ("Lab" Table 4).

Table 4. Variance components for the Local and Standardized protocols in Stage 3 for the 2.5 mg/kg Clozapine treatment (model 1).

Parameter	Local	Standardized
Lab	1.163 (26.38%)	0.436 (12.97%)
Residual	3.247 (73.62%)	2.926 (87.03%)
Total	4.410 (100.00%)	3.362 (100.00%)

Finally, an across-stage comparison (model 3) was made between the control condition of the harmonized protocol (standardized cohort) in stage 3 and the control condition of the same protocol from stage 2. The different stages yielded similar variance compo-

nents (Table 5, model 1). The variance introduced by using the Standardized protocol in different stages with different vehicles was <1% (Table 6).

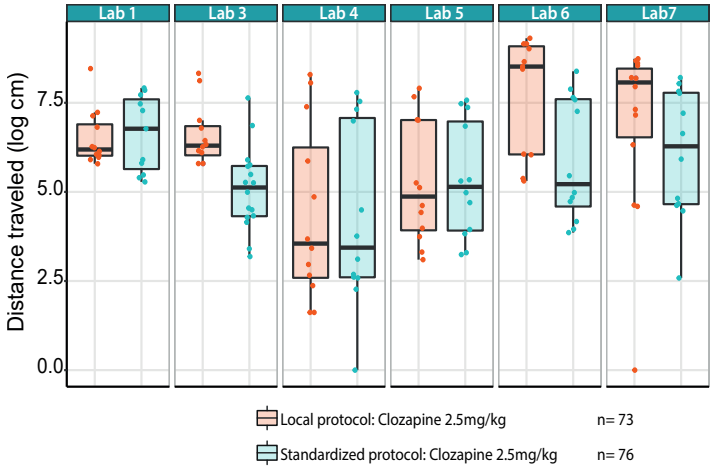


Fig 5. Distance traveled across sites following Clozapine administration (2.5 mg/kg) after implementation of the Local protocol (purple) and Standardized protocol (teal) in stage 3. Data underlying this figure can be found in https://osf.io/8f6yr/, Stage 3.

Table 5. Variance components for the Standardized protocol at its control conditions across stages 2 and 3 (model 1).

Parameter	Standardized (Stage 2)	Standardized (Stage 3)
Lab	0.016 (29.90%)	0.022 (31.01%)
Residual	0.038 (70.10%)	0.048 (68.99%)
Total	0.055 (100.00%)	0.069 (100.00%)

Table 6. Comparison across stages 2 and 3 of the standardized protocol for the control condition (model 3).

Parameter	Std. Deviation	Variance Estimate	%
Stage	0.021	<0.001	0.75
Lab	0.128	0.017	26.90
Residual	0.211	0.044	72.35
Total	0.361	0.061	100.00

Lastly, the impact of sex as a blocking factor was explored across laboratories as a fixed effect (Supplementary Tables B and C in S3 Supplementary Stage). This analysis revealed that sex did not affect the outcome measure as it did not explain the variance of the data.

DISCUSSION

Overall, our study shows that harmonization of experimental protocols across sites reduced the outcome variability across laboratories compared to site-specific versions of the protocol (*i.e.,* local protocol). Moreover, we found that sex did not affect the results and that illumination of the test arena, and time of testing relative to the light-dark cycle were not suitable factors to systematically introduce variation in the results of an open field test in C57BL/6 mice. Regarding the time of testing, we could speculate that the treatment effect had such a strong effect on the outcome variable that there was no room for the time variable to further affect the outcome. Another possible explanation is that this environmental factor does not have a strong influence on the particular outcome tested with the current experimental setup (e.g., the drug and dose used).

The present study showed that between-lab variation is rather large when lab-specific protocols are followed (*e.g.,* local protocol) and although it was reduced by protocol harmonization, it remained considerable. This corroborates earlier findings (14) that site-specific variation in conditions produces between-lab variability that cannot be neutralized by protocol harmonization across sites. This in turn affects the replicability of study outcomes.

Although the standardized protocol successfully produced replicable results across laboratories, the sensitivity to detect drug treatment effects can still be improved as not all sites found a significant drug treatment effect in stage 3 for the lowest dose (Fig 4; right panel). The choice of the two doses tested in stage 3 was based on a literature review performed by one of the partners where the higher dose had a robust effect while the lower dose showed conflicting results. It seems possible that the discrepancy between the sites is due to inherent differences between laboratories that were heightened by the stringent local standardization. It was suggested that a way around this would be to introduce systematic variation within sites, hoping this will account for the variance between sites and test the same drug treatments (17,23).

To test this hypothesis, we introduced systematic variation to the standardized protocol. Contrary to our expectation, this heterogenized cohort did not increase the overall variability, and neither did it decrease the between laboratory variability in outcomes when compared to standardized alone. The overall outcome of the results did not change (*i.e.,* similar drug treatment effects were obtained following the heterogenized and standardized cohorts). Therefore, we could not confirm that diversifying the environmental conditions further reduces the variability across laboratories. The current selection of 'heterogenizing' factors was rather limited by the feasibility to diversify them across all

labs. Further factors, for example, genotypic variation of the study sample, should be considered for future studies as they may have stronger power to introduce within-study variability than environmental variability as seen in other disciplines (24). A recent initiative that could prove helpful for identifying heterogenization factors is the Platform for the Exchange of Experimental Research Standards (PEERS) developed to rate the factors and variables most likely to influence experimental outcomes (25).

Moreover, the standardized protocol showed to be robust to the introduction of animals of both sexes in stage 3. Sex did not increase the variability of results across sites compared to the standardized protocol (Table C in S3 Supplementary Stage) and did not account for the variance in the data. In this case, sex may be included without a need to increase the sample size. However, sex should always be included as a biological variable in biomedical research for reasons of inclusion, regardless of its effect on the results (23). While the harmonization of a standardized protocol across laboratories decreased the overall variability of results compared to when each laboratory followed its own local protocol, the question arises whether these results, although replicable across the participating laboratories, could be further generalized to other laboratories outside the present study. Assuming that the participating laboratories are a representative random sample of laboratories doing phenotyping studies, we could say our results can be extrapolated to other laboratories; however, caution must be taken as the participating labs were all highly interested in data quality and results replicability . This fact might have biased the current sample.

To be able to extrapolate an experimental result to other conditions or populations (i.e., have a broad inference space) the study population has to be representative of the desired target population. Our finding that systematically introducing additional factors (illumination and time of testing in stage 2 and sex in stage 3) did not affect the overall variation, shows that diversifying a study population and its environment does not necessarily lead to more "noisy" experimental outcomes, but allows to broaden the inference space and increase the external validity of the results and thus their generalizability (26). This supports diversifying environmental factors that (i) are not tightly linked with the outcome measure or (ii) are not directly involved in the research question as a means to increase the robustness of results. On the other hand, it is necessary to continue exploring the effects of protocol harmonization in results variability since our results suggest that although harmonizing protocols across laboratories reduced between-lab variation, the laboratory factor explains most of the variance, meaning that standardizing is not enough.

CONCLUSION

Altogether, we can say that both harmonized (*i.e.,* standardized and heterogenized) open field protocols consistently and significantly reduced the between-lab variability of the behavioral outcome. In addition, the protocols resulted in consistent treatment effects across laboratories that were also replicable within laboratories across the different stages. The replicability of results within and between laboratories in the present study highlights the impact of study-specific variation in between-lab variability, and the importance of transparent and fine-grained reporting of methodologies, and research protocols. It also shows that it is possible to diversify the study sample by incorporating blocking factors like sex or introducing systematic heterogenization of conditions without the need to increase the overall sample size.

MATERIALS AND METHODS

General outline

The experiment compared the variability of open field activity in mice after pharmacological treatment across seven laboratories in Europe, Israel and the United States, including academic and industry sites. All sites concurrently followed a 3-stage approach wherein different experimental protocols were implemented with the aims to (i) assess the contribution of laboratory-specific (local) protocols to between-lab variability compared to a fully harmonized protocol, and (ii) compare a standardized cohort with a heterogenized cohort to assess whether increased diversity enhances external validity, resulting in enhanced replicability.

The selection of the open field test was based on frequent use in the field of biomedical and neuroscience research for the assessment of behavior, and specifically for the measurement of locomotor activity levels. Because the purpose of this project was to develop a mechanism for ensuring the concordance of generated data, we decided to focus on one of the simplest yet ubiquitous aspects of behavior, namely locomotion with distance traveled being the primary outcome measure. The ex-ante study protocols per site and stage, and raw data are publicly available in the OSF repository (DOI: 10.17605/OSF.IO/8F6YR).

Laboratory sites and ethical statements

All animal procedures were carried out following the regulations of Directive 2010/63/EU or the Association for Assessment and Accreditation of Laboratory Animal Care

and following the recommendations of the Guide for the Care and Use of Laboratory Animals. The individual ethical committee for each institution can be found in Table 7.

Table 7. Ethical approval committees for each of the laboratories involved

Laboratory	Ethical approval body
GELIFES (Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands)	Animal Welfare Body of the University of Groningen and the National Central Committee for scientific procedures on animals (CCD)
LMU (Ludwig-Maximilians-Universitaet Muenchen, Muenchen, Germany)	Government of Upper Bavaria (reference number ROB-55.2-2532.Vet_02-18-45)
Organon (Organon RD, Turku, Finland)	Project Authorization Board in the Regional State Administrative Agency for Southern Finland
PsychoGenics Inc. (New Jersey, USA)	Institutional Animal Care and Used Committee (IACUC #271)
Sylics (Synaptologics BV, Amsterdam, The Netherlands)	Animal Welfare Body of the VU University Amsterdam and the National Central Committee for scientific procedures on animals (CCD)
Teva Pharmaceuticals (Tel Aviv, Israel)	Animal welfare council of the Ministry of Health of Israel (internal committee request #715).
UBERN (Universitaet Bern, Bern, Switzerland)	Cantonal Veterinary Office of the Canton of Bern: License number BE 18/18

In addition, all sites ensured detailed recording of experimental method and procedure according to "The ARRIVE guidelines Animal Research: Reporting In Vivo Experiments" (27) and adhered to the EQIPD key principles for guiding the design, conduct and analysis of preclinical efficacy and safety research (22).

Experimental design

a. Animals

All experiments were performed with C57BL/6J mice. The details regarding the age, sex, and origin of the animals are summarized in Table 8 as they differed across stages. Likewise, and as part of the three-stage experimental approach animal numbers, housing, husbandry and experimental conditions also varied across stages (see Table 8). Note that the animal numbers in Table 8 represent the result of the power calculation; however, some sites included more animals to, for example, even out the total number of animals per group. Thus, the number of observations differs from the one required on Table 8. The animals used in each protocol were experimentally naïve and came from independent batches at all sites. The use of different animal providers among laboratories served as a representation of common differences between laboratories and study populations to test the performance of the different protocols.

b. Design

The readouts of distance traveled in the open field were collected during three consecutive stages with around one year apart, in seven different laboratories. In stage 1, all sites performed the study with minimal alignment (strain, age, drug treatment, vehicle, primary outcome measure, and test duration) using their 'in-house' standard operation procedures (SOP) under the local conditions at each site (*e.g.,* light intensity, arena size, husbandry conditions, etc.); this is referred to as local protocol and was intended as a baseline measurement of the variability between a 'random' sample of laboratories. In stage 2, specific husbandry and experimental conditions were harmonized across laboratories. In addition, besides a standardized cohort of the harmonized protocol, with all factors standardized within laboratories, a heterogenized cohort of the harmonized protocol was used; this cohort aimed to increase within-site variability of the data by systematically varying 2 selected 'heterogenization' factors using a 2x2 factorial design. Finally, the stage 3 goal was to challenge the sensitivity of the standardized cohort of the harmonized protocol from stage 2 by using the same protocol but with a different drug treatment than previously used (Table 8). The local protocol used in stage 1 was replicated in stages 2 and 3 as a control condition for the harmonized protocols, and to assess the replicability of results obtained with the local protocol across stages within each of the laboratories.

An a-priori power analysis (G*Power v3.1.9.2) was performed based on effect sizes estimated from literature (25) and a previous study with the NMDA receptor antagonist MK-801 performed by one of the partners. The study was powered so that each treatment for each sex and each laboratory is treated as a stand-alone test. Alpha was set to 0.05 and power to 0.9 for a t-test for means-difference for two independent means. The calculated effect sizes were considered large. To evaluate the required n numbers in the case of a 'medium' effect size, the effect size was halved and corresponding n numbers were again estimated. The required n numbers for 'MK´ 801 0.3 mg/kg' and '3 mg/kg diazepam' were considered too low and were therefore increased to 6. Recommendations of a minimum number of animals to enroll per drug and dose treatment are shown in Table 8. This power calculation was used for the stage 1 protocol. The number of animals used per stage was adapted according to the compounds used and the number of animals available in the animal use licenses.

c. Pharmacological compounds

In each of the stages, the spontaneous locomotor activity after acute administration of a compound was compared across treatment groups and/or dosages. While the compounds and dosage used varied across stages (Table 8), the pretreatment time was kept constant with administration 30 minutes before the start of the test. Diazepam (Duchefa

Biochemie, BUFA, Roche, Merck, Sigma Aldrich, TEVA) was used in stage 1 together with MK-801 (Sigma Aldrich); the latter was also used in stage 2. Clozapine (Sigma Aldrich cat# C6305) was administered in stage 3. Data from the different drug treatments were compared with the control group that received the respective vehicle (*i.e.,* drug dose= 0 mg/kg). The vehicle was different across treatments according to the compound solubility (see Table 8).

The dose range of each compound was selected according to the goal of each stage as follows. Stage 1: At the localization stage, we aimed for a dosage with a strong effect and a dosage with a medium effect, assuming replicability would be lower with a subtler effect. Doses were based on previous data collected from one of the partners. However, from this, we expected 0.3mg/kg to have a stronger effect (hence the smaller sample size), but it turned out that it was the other way round. Therefore, we used the smaller dose in stage 2. Stage 2: At stage 2, the focus was on comparing localization with harmonization (primary aim) and standardization with heterogenization (secondary aim); because this increased the number of treatment groups considerably, we limited the study to a single dose against saline control. For the same reason, we limited the study to a single sex (choosing females to minimize the risk of injury and aggression, which is more frequent in males). Thus, stage two is a kind of proof-of-principle study to inform stage 3. Stage 3: Similar to stage 1, we wanted a dose with a strong effect and a dose with a weaker effect

d. Protocols

Local protocol: a predefined minimum set of requirements were aligned across sites (Table 8). Variables not addressed in these minimum requirements were to be handled according to the SOP of each site. All variables, both aligned and non-aligned, were reported *post hoc* following stage 1 tests to generate an inventory of the different environmental variables that may have influenced the between laboratory variability. This protocol represents the most common scenario in preclinical biomedical animal research, where independent studies are standardized within laboratories but conditions and procedures vary between laboratories.

The local protocol was replicated in all stages to test the replicability of results within each laboratory, and as the control protocol to compare the other protocols at each stage. Therefore, the number of animals, drug treatments and vehicles differed across stages according to Table 8.

Harmonized protocol – standardized cohort: from the local protocol inventories, variables that differed between sites were identified and chosen based on their biologi-

cal relevance and feasibility to be modified at all sites; these were further harmonized across sites to test whether the variability of effect sizes across sites observed in the local stage could be reduced by controlling for these variables. This cohort aimed to assess how much the lab-specific differences in the standardized protocol contribute to between-lab variation. As indicated in Table 8, this protocol was used in stages 2 and 3; the treatment, vehicle and treatment dosage differed between stages as well as the inclusion of male mice in stage 3.

Harmonized protocol - Heterogenized cohort: this protocol was identical to the standardized cohort, with the exception that two factors were systematically varied within sites to account for the variability between sites, namely light intensity in the experimental arena was set to either dim (20 Lux) or bright (80 Lux) and the window time of testing concerning the light-dark phase was varied between early (2-4 hours after light on) or late (8-10 hours after lights on).

The standardized and heterogenized cohorts of the harmonized protocol were tested in parallel, to assess whether simple heterogenization of environmental factors would further reduce the between laboratory variability compared to the local protocol.

Experimental procedures

a. Behavioral assessment

The Open Field (OF) test was used throughout the study to measure the effect of different pharmacological compounds (Table 8) on locomotor activity. Mice were placed in the center of an empty open arena and horizontal activity was recorded for 15 minutes, afterwards, the animal was taken back to its home cage. The outcome measure was the total distance traveled in the OF arena. Details on the open field arenas, recording and scoring methods are summarized in Table 8.

Table 8. Variables and corresponding values across the different stages and their respective protocol(s) followed by all sites.

FACTOR	STAGE 1	STAGE 2		STAGE 3
	Local	**Standardized**	**Heterogenized**	**Standardized**
Rearing and housing				
Experimental animals				
Sex	Female and male	Female	Female	Female and male
Strain	C57BL/6J	C57BL/6J	C57BL/6J	C57BL/6J
Age	8-10 weeks	9 weeks	9 weeks	9 weeks

Table 8. Variables and corresponding values across the different stages and their respective protocol(s) followed by all sites. (continued)

FACTOR	STAGE 1	STAGE 2		STAGE 3
	Local	**Standardized**	**Heterogenized**	**Standardized**
Provider	Variable (In-house, Janvier Lab, Charles River Lab [Germany and France], Envigo [NL and Jerusalem], Jackson Lab)	Variable (In-house, Janvier Lab, Charles River Lab [Germany and France], Envigo [NL and Jerusalem], Jackson Lab)	Variable (In-house, Janvier Lab, Charles River Lab [Germany and France], Envigo [NL and Jerusalem], Jackson Lab)	Variable (In-house, Janvier Lab, Charles River Lab [Germany and France], Envigo [NL and Jerusalem], Jackson Lab)
Housing				
Animals per cage	2-5 By sex	3	3	2 By sex
Same sex cage mates	Yes	Yes	Yes	Yes
Cage size	Makrolon I, II L or III	Makrolon III	Makrolon III	Makrolon III
Environmental enrichment type	Variable (e.g., nesting material and shelter or tube and nesting material)	Only 1 type of enrichment (e.g., nesting material or tunnel or shelter)	Only 1 type of enrichment (e.g., nesting material or tunnel or shelter	Only 1 type of enrichment (e.g., nesting material or tunnel or shelter
Husbandry				
Handling method	Tail or cupped with gloved hands	Tail with gloved hands	Tail with gloved hands	Tail with gloved hands
Handling frequency	1-2 times x week	1 time x week	1 time x week	1 time x week
Behavioral testing				
Experimenter gender	Variable	Female	Female	Female
Number of handlers	Multiple	Single/Two[#] (1 person doing all injections and/or 1 performing the experiment)	Single/Two[#] (1 person doing all injections and/ or 1performing the experiment)	Single/Two[#] (1 person doing all injections and/or 1performing the experiment)
Acclimation to experimental room	Variable (0-60 min)	60 min	60 min	60 min
Acquisition method	IR beam breaks or[#] Video tracking	IR beam breaks or[#] Video tracking	IR beam breaks or[#] Video tracking	IR beam breaks or[#] Video tracking
Test arena cleaning method	Variable	Tap water	Tap water	Tap water
Drug treatment tested	Diazepam (Dz) & MK-801 (MK)	Mk-801	Mk-801	Clozapine (Clz)
Drug treatment dosage	Dz: 3 mg/kg MK: 0.2 & 0.3 mg/kg	0.2 mg/kg	0.2 mg/kg	1 & 2.5 mg/kg
Injection volume & route	10 mL/kg i.p.	10 mL/kg i.p.	10 mL/kg i.p.	10 mL/kg i.p.

Table 8. Variables and corresponding values across the different stages and their respective protocol(s) followed by all sites. (continued)

FACTOR	STAGE 1	STAGE 2		STAGE 3
	Local	**Standardized**	**Heterogenized**	**Standardized**
Vehicle	MK: Saline DZ:40% propylene glycol + 10% alcohol + 50% Saline	Saline	Saline	Ultrapure water
Experimental groups	5	2	2	3
Sample sizes	Saline: 28 (14 Females) MK 0.2 mg/kg: 28 (14F) MK 0.3mg/kg: 12 (6F) Vehicle Dz: 12 (6F) Dz: 12 (6F)	Saline: 12 MK-801: 12	Saline: 12 MK-801: 12	Ultrapure water: 16 (8F) Clz 1 mg/kg: 16 (8F) Clz 2.5 mg/kg: 12 (6F)
Treatment assignment	Variable: random number generator or pick randomly from cage	Block-randomized or balanced across cages*	Block-randomized or balanced across cages*	Block-randomized or balanced across cages*
Blinded performance and scoring		Yes	Yes	Yes
Test duration	At least 15 min	15 min	15 min	15 min
Test phase	Light	Light: 4-8 hours after lights ON	Light *Early:* 2-4 hours OR *Late:* 8-10 hours after lights ON	Light: 4-8 hours after lights ON
Outcome variable	Distance traveled	Distance traveled	Distance traveled	Distance traveled
Experimental unit	mouse	cage	cage	cage
Experimental Setup				
OF arena shape	Circular or[#] square	Square (28 x 28 cm)	Square (28 x 28 cm)	Square (28 x 28 cm)
OF arena color	White, gray or black	White	White	White
OF arena light intensity	20-350 Lux	50 Lux	Dim: 20 Lux OR Bright: 80 Lux	50 Lux

*All animals in a cage received the same treatment and were tested in parallel. The reasoning behind this was to avoid social facilitation from 'agitated' mice after MK-801 injection influencing control/vehicle mice during the 30-min wait between the injection and the test ; cages and/or animals were block randomized according to the cage location using Blindr tool developed by the VU Amsterdam (https://github.com/jhuebotter/Blindr). [#]According to the availability at each site; for details see Table S1 of the Stage 2 protocol available in OSF.

Animal handling and drug administration, scoring, and analyses were performed blinded to the treatments unless stated otherwise. This means that the person handling and dosing the animals was not aware of the allocation of animals into experimental

groups nor about which of the treatments was being administered. Animals were block-randomized into groups by various methods (Blindr; random number generator from Mathematica v11, Wolfram Inc. ; R-script provided by one of the partners or developed in-house as part of their Data Management software) except for site 5 which used even distribution of animals into groups. The outcome measure was the total distance traveled for 15 minutes in the Open Field.

There were no predetermined exclusion criteria unless animals presented health issues. However, some sites were able to add animals to, for example, even out the number of animals in each treatment group. Therefore, the raw data has more animals than the ones mentioned in Table 8 for some of the sites.

b. Data management

Once data were acquired, each site transferred its raw data and metadata to an Excel structure shared across sites, which was then shared for centralized analysis. Each site was responsible for checking the soundness of the data (*i.e.,* quality-check for the video length, accurate scoring, correct group coding, etc.). Averaged data from each treatment group per site across stages can be found in the S1, S2, S3 Supplementary Stage files.

c. Statistical methods

Before the analysis a log transformation was performed to the outcome variable (*i.e.,* total distance traveled) because data are naturally bounded between 0 and + infinity. The log transformation changes the bound and sets it between -infinity and +infinity which is more aligned with the assumptions of linear modeling. Moreover, we are sure that the model accounts for those natural bounds when estimating the effects. Otherwise, it could be that some effects have a 95% CI lower bound lower than 0 which would be uninformative given the outcome variable analyzed.

<u>Within stages and protocols analysis</u>

To study the lab-to-lab variation by stage and protocol, two types of models were used. The first one explored the differences in dosing effects by the laboratory. This analysis reflects a situation where each laboratory would perform the comparisons internally and aims to highlight the variability of estimated differences between laboratories. It was expected that the Harmonized protocol provides more consistent results than the Local protocol. A simple linear regression was fitted to the natural logarithm transform of total distance travelled by laboratory with drug treatment as a unique fixed effect:

$$Y_{id} = \beta_0 + \beta_d \times dose_d + \varepsilon_{id} \text{ (model 1)}$$

where Y_{id} is the natural logarithm transform of total distance travelled *i* for drug treatment *d*,

β_0 is the intercept of the model (the expected Y_{id} for drug treatment *d* of reference),

β_d is the effect of drug treatment *d* on Y_{id} (the expected change in Y_{id} when drug treatment *d* is considered),

ε_{id} is the random error associated with Y_{id}: $\varepsilon_{id} \sim N(0, {\sigma^2}_\varepsilon)$ where σ^2_ε is the residual or biological variance

The drug treatment effects and their contrasts were estimated using the R package emmeans.

Note that for the standardized protocol in stage 3, Y_{id} is the natural logarithm transform of total distance travelled plus 1 because of some zero values for which the natural logarithm would not be defined. Moreover, also for the standardized protocol in stage 3, huge discrepancies were observed between variances of drug treatment. Hence, one variance per drug treatment was modelled instead of one pooled variance: Y_{id}: $\varepsilon_{id} \sim N(0, {\sigma^2}_\varepsilon)$. This specific model with individual variance per drug treatment was fitted with the R package glmmTMB.

The second model explores the differences in dosing effects overall laboratories accounting for the lab-to-lab variability. It aimed to directly estimate the variance associated with differences between laboratories and assess the percentage of total variance it represented. It was expected that the Harmonized protocol provides lower lab-to-lab variance than the Local protocol while having similar residual variances. A linear mixed model was fitted to the natural logarithm transform of total distance travelled with drug treatment as a unique fixed effect and laboratory as well as the interaction between laboratory and drug treatment as random effects:

$$Y_{idl} = \beta_0 + \beta_d \times dose_d + b_l + d_{dl} + \varepsilon_{idl} \text{ (model 2)}$$

where Y_{idl} is the natural logarithm transform of total distance travelled *i* for drug treatment *d* and lab *l*,

β_0 is the intercept of the model (the expected Y_{idl} for drug treatment *d* of reference),

β_d is the effect of drug treatment *d* on Y_{idl} (the expected change in Y_{idl} when drug treatment *d* is considered),

b_l = the random intercept of laboratory l: $b_l \sim N(0, \sigma^2_b)$,

d_{dl} = the random intercept of drug treatment *d* and laboratory l: $d_{dl} \sim N(0, \sigma^2_d)$ and,

ε_{idl} is the random error associated with Y_{idl}: $\varepsilon_{idl} \sim N(0, {\sigma^2}_\varepsilon)$.

${\sigma^2}_b, {\sigma^2}_d$ and ${\sigma^2}_\varepsilon$ are referred as the variance components in this model. In linear mixed model, variance is decomposed in several terms of interest to understand which ones

are the main source of variability in the data. In this specific case, $\sigma^2{}_b$ is the lab-to-lab variability, $\sigma^2{}_d$ is the variability in differences between drug doses observed between lab and $\sigma^2{}_\varepsilon$ is the residual or biological variability.

The model was fitted with R package lmer. The drug treatment effects and their contrasts were estimated using the R package emmeans. Note that for the standardized protocol in stage 3, the same modifications were applied as for the simple linear model.

Between stages analysis

To study the stage-to-stage variation for one protocol, two types of approaches were used. The first one compared the within-stage results by stage for common dosing groups. It assesses if effects observed in laboratories and if the variance components are similar from one stage to another. This would indicate that the results are replicable. Local protocols of stages 1 and 2 are compared using the common control and MK-801-0.2mg/kg groups, whereas Harmonized protocols (standardized cohort) stages 2 and 3 are compared using the common 2.5mg/kg clozapine treatment. Note that the models presented simplify for the standardized cohort of the Harmonized protocol because there is only one dose.

The second model explored the ffectts over all laboratories accounting for lab-to-lab and stage-to-stage variability. It aimed to estimate the proportion of the total variance attributable to between-stage differences. A linear mixed model was fitted to the natural logarithm transform of the total distance travelled with drug treatment as unique fixed effect and laboratory as well as interaction between laboratory and drug treatment, stage and the interaction between stage and drug treatment as random effects:

$$Y_{idls} = \beta_0 + \beta_d \times dose_d + b_l + d_{dl} + s_s + g_{ds} + \varepsilon_{idls} \quad \text{(model 3)}$$

where Y_{idls} is the natural logarithm transform of total distance travelled *i* for drug treatment *d*,
lab *l* and stage *s*,
β_0 is the intercept of the model (the expected Y_{idls} for drug treatment *d* of reference),
β_d is the effect of drug treatment *d* on Y_{idls} (the expected change in Y_{idls} when drug treatment *d* is considered),
b_l = the random intercept of laboratory l: $b_l \sim N(0, \sigma_b^2)$,
d_{dl} = the random intercept of drug treatment and laboratory l: $d_{dl} \sim N(0, \sigma_d^2)$,
s_s is the random intercept of stage s: $s_s \sim N(0, \sigma_s^2)$,
g_{ds} is the random intercept of drug treatment *d* and stage s: $g_{ds} \sim N(0, \sigma_g^2)$ and,

ε_{idls} is the random error associated with Y_{idls}: $\varepsilon_{idls} \sim N(0, {\sigma^2}_{\varepsilon})$, where ${\sigma^2}_{\varepsilon}$ is the residual or biological variance.

Influence of external factors

Additional models were performed to study the effects of heterogeneous factors introduced for the Harmonized protocol – Heterogenized cohort (the light intensity and the time of testing) in stage 2 and the blocking factor for the Harmonized protocol in stage 3 (sex) on the data. Two approaches were considered, first a by-laboratory analysis, then an across laboratory analysis, both by factor of interest. The models were based on the ones used in the within stages and protocols analysis. Two fixed effects were added each time, the factor and the interaction between the factor and the drug treatment. The models were fitted with R package lmer. Statistical significance of those effects was tested with F-tests (Type III) using the R package lmerTest. The different effects and their contrasts were estimated using the R package emmeans.

The boxplots were computed with the raw data that can be found in https://osf.io/8f6yr/, while the treatment effect differences are reported in the S1, S2, and S3 Supplementary Stage files. See each figure for specifics. The analysis codes can be found in the OSF repository (DOI: 10.17605/OSF.IO/8F6YR).

ACKNOWLEDGEMENTS

We would like to thank Eva-Lotta von Rüden and Sarah Glisic for their contribution to the Muenchen site.

This publication reflects only the authors' view and the Innovative Medicines Initiative 2 Joint Undertaking is not responsible for any use that may be made of the information it contains.

REFERENCES

1. Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, et al. Investigating the replicability of preclinical cancer biology. Pasqualini R, Franco E, editors. eLife. 2021 Dec 7;10:e71601.
2. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Science Translational Medicine. 2016 Jun 1;8(341):341ps12-341ps12.
3. Bishop D. Rein in the four horsemen of irreproducibility. Nature. 2019 Apr;568(7753):435–435.
4. Giles J. Animal experiments under fire for poor design. Nature. 2006 Dec 1;444(7122):981–981.
5. Kafkafi N, Agassi J, Chesler EJ, Crabbe JC, Crusio WE, Eilam D, et al. Reproducibility and replicability of rodent phenotyping in preclinical studies. Neurosci Biobehav Rev. 2018 Apr;87:218–32.
6. Loken E, Gelman A. Measurement error and the replication crisis. Science. 2017 Feb 10;355(6325):584–5.
7. Steward O, Balice-Gordon R. Rigor or Mortis: Best Practices for Preclinical Research in Neuroscience. Neuron. 2014 Nov 5;84(3):572–81.
8. Fanelli D. Is science really facing a reproducibility crisis, and do we need it to? Proceedings of the National Academy of Sciences. 2018 Mar 13;115(11):2628–31.
9. Richter SH, Garner JP, Würbel H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nat Methods. 2009 Apr;6(4):257–61.
10. Würbel H. Behaviour and the standardization fallacy. Nat Genet. 2000 Nov;26(3):263.
11. Voelkl B, Würbel H. A reaction norm perspective on reproducibility. Theory Biosci. 2021 Jun;140(2):169–76.
12. Schlichting, C.D., Pigliucci, M. Phenotypic evolution: a reaction norm perspective. Sinauer Associates; 1998.
13. Brown SDM, Moore MW. The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping. Mamm Genome. 2012 Oct;23(0):632–40.
14. Crabbe JC, Wahlsten D, Dudek BC. Genetics of mouse behavior: interactions with laboratory environment. Science. 1999 Jun 4;284(5420):1670–2.
15. Arroyo-Araujo M, Graf R, Maco M, van Dam E, Schenker E, Drinkenburg W, et al. Reproducibility via coordinated standardization: a multi-center study in a S hank 2 genetic rat model for Autism Spectrum Disorders. Scientific Reports. 2019 Aug 12;9(1):1–10.
16. Wahlsten D, Metten P, Phillips TJ, Boehm SL, Burkhart-Kasch S, Dorow J, et al. Different data from different labs: lessons from studies of gene-environment interaction. J Neurobiol. 2003 Jan;54(1):283–311.
17. Richter SH, Garner JP, Auer C, Kunert J, Würbel H. Systematic variation improves reproducibility of animal experiments. Nature Methods. 2010 Mar;7(3):167–8.
18. Voelkl B, Vogt L, Sena ES, Würbel H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLOS Biology. 2018 Feb 22;16(2):e2003693.
19. Bodden C, Kortzfleisch VT von, Karwinkel F, Kaiser S, Sachser N, Richter SH. Heterogenising study samples across testing time improves reproducibility of behavioural data. Scientific Reports. 2019 Jun 3;9(1):8247.
20. Bailoo JD, Voelkl B, Varholick J, Novak J, Murphy E, Rosso M, et al. Effects of weaning age and housing conditions on phenotypic differences in mice. Sci Rep. 2020 Dec;10(1):11684.
21. Karp NA, Wilson Z, Stalker E, Mooney L, Lazic SE, Zhang B, et al. A multi-batch design to deliver robust estimates of efficacy and reduce animal use – a syngeneic tumour case study. Sci Rep. 2020 Apr 10;10(1):6178.

22. Bespalov A, Bernard R, Gilis A, Gerlach B, Guillen J, Castagne V, et al. Introduction to the EQIPD quality system. Zaidi M, editor. eLife. 2021 May 24;10:e63294.
23. Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, Jaric I, et al. Reproducibility of animal research in light of biological variation. Nat Rev Neurosci. 2020 Jul;21(7):384–93.
24. Milcu A, Puga-Freitas R, Ellison AM, Blouin M, Scheu S, Freschet GT, et al. Genotypic variability enhances the reproducibility of an ecological study. Nat Ecol Evol. 2018 Feb;2(2):279–87.
25. Sil A, Bespalov A, Dalla C, Ferland-Beckham C, Herremans A, Karantzalos K, et al. PEERS — An Open Science "Platform for the Exchange of Experimental Research Standards" in Biomedicine. Frontiers in Behavioral Neuroscience. 2021;15:256.
26. Usui T, Macleod MR, McCann SK, Senior AM, Nakagawa S. Meta-analysis of variation suggests that embracing variability improves both replicability and generalizability in preclinical research. PLOS Biology. 2021 May 19;19(5):e3001009.
27. Sert NP du, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, et al. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLOS Biology. 2020 Jul 14;18(7):e3000411.

SUPPORTING INFORMATION

S1 Supplementary Stage. Supplementary tables with raw data and statistical results of stage 1.

S2 Supplementary Stage. Supplementary tables with raw data and statistical results of stage 2.

S3 Supplementary Stage. Supplementary tables with raw data and statistical results of stage 3.

S1 Methods. Video analysis settings and husbandry details

S1 Protocol. Supplementary Protocol per site for Stage 1 with the minimum requirements asked.

Translational validity and methodological underreporting in animal research: a systematic review and meta-analysis of the Fragile X syndrome (Fmr1 KO) rodent model

Renate Kat[a,c], **María Arroyo-Araujo**[a,c], Rob B.M. de Vries[b], Marthe A. Koopmans[a], Sietse F. de Boer[a], Martien J.H. Kas[a,d]

[a]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Nijenborgh 7, 9747 AG, Groningen, The Netherlands,
[b]SYRCLE, Department for Health Evidence, Radboud Institute for Health Sciences, Radboud University Medical Centre, Geert Groteplein Zuid 21, 6525 EZ, Nijmegen, The Netherlands,
[c]These authors contributed equally to this work

Published in Neuroscience and Biobehavioural Reviews (2022, 139:104722)

ABSTRACT

Predictive models are essential for advancing knowledge of brain disorders. High variation in study outcomes hampers progress. To address the validity of predictive models, we performed a systematic review and meta-analysis on behavioural phenotypes of the knock-out rodent model for Fragile X syndrome according to the PRISMA reporting guidelines. In addition, factors accountable for the heterogeneity between findings were analyzed. The knock-out model showed good translational validity and replicability for hyperactivity, cognitive and seizure phenotypes. Despite low replicability, translational validity was also found for social behaviour and sensory sensitivity, but not for attention, aggression and cognitive flexibility. Anxiety, acoustic startle and prepulse inhibition phenotypes, despite low replicability, were opposite to patient symptomatology. Subgroup analyses for experimental factors moderately explain the low replicability, these analyses were hindered by under-reporting of methodologies and environmental conditions. Together, the model has translational validity for most clinical phenotypes, but caution must be taken due to low effect sizes and high inter-study variability. These findings should be considered in view of other rodent models in preclinical research.

Keywords: Autism spectrum disorder, mouse models, preclinical data quality

INTRODUCTION

The Fragile X Syndrome (FXS) is a common inherited form of intellectual disability and one of the most prominent genetic causes of syndromic autistic spectrum disorders (ASD; Kidd et al., 2014). FXS is caused by a CGG repeat mutation on the X chromosome containing the *FMR1* gene, causing a deficiency of the resultant protein (Verkerk et al., 1991). The *FMR1* gene codes for the RNA-binding protein fragile X mental retardation protein (FMRP), which binding targets include several synaptic proteins essential for proper neurotransmission and neuronal structure, affecting multiple neuronal pathways. Individuals carrying the full FMRP mutation typically display intellectual disabilities, seizures, attention deficits, increased anxiety, hyperarousal to stimuli, and macroorchidism together with autistic-like features. Of the FXS patients, 30% meet the criteria for ASD diagnosis (Bailey et al., 1998; Baumgardner et al., 1995; Hagerman et al., 1986; Hersh et al., 2011), but up to 90% of patients show some of the symptoms of ASD (Hagerman et al., 1986). In general, females display milder symptoms than males.

The FMRP lack of expression was successfully reproduced in mice to generate an animal model to study. The most frequently used *Fmr1* KO mouse model came from the Dutch-Belgian Fragile X Consortium (The Dutch-Belgian Fragile X Consortium et al., 1994) and does not produce FMRP because of a disruption in the *FMR1* DNA sequence with an insertion in exon 5. Still, it has a detectable level of *FMR1* mRNA (Kazdoba et al., 2014). A second-generation KO model (KO2) was later developed which no longer has *Fmr1* mRNA present (Mientjes et al., 2006). The majority of research on these models have focussed on the affected molecular pathophysiological pathways, like increased immature spine densities and GABA-ergic deficits, which have recently been reviewed elsewhere (Dionne and Corbin, 2021; Telias, 2019). Additionally, a large body of literature has reported on the behavioural abnormalities of these models. These mouse models, as well as some KO rat models, have been reported to recapitulate several phenotypic features seen in patients such as cognitive deficits, social anxiety, reduced social interaction, repetitive behaviours and hyperactivity. However, there is a considerable number of contrasting findings in the literature. For example, while many papers report inhibitory avoidance cognitive deficits in *Fmr1* KO mice (Ding et al., 2020, 2014; Li et al., 2020; Qin et al., 2015; Saré et al., 2016) other studies found no difference between KO and wildtype (WT) mice using the same task (Liao et al., 2018; Melancia and Trezza, 2018; Saré et al., 2019, 2018; The Dutch-Belgian Fragile X Consortium et al., 1994). These discrepancies are also found for tasks that test for recognition memory, social discrimination, and spatial memory and, more importantly, for tasks that measure the core symptomatic features of ASD such as tasks that evaluate social behaviour, repetitive behaviour, communication, and anxiety. Recently in our lab, the mouse *Fmr1* KO model was tested in a behavioural bat-

tery to assess repetitive and social behaviours. To our surprise, no apparent phenotype was found, also contrasting with the behavioural repertoire seen in patients and sometimes found in the preclinical literature.

It has been suggested that differences in methodological approaches and diverse research practices can impact the behavioural outcome measures, which may partly explain the contrasting literature. However, preclinical research has also shown a lack of transparency of reporting as well as the use of inappropriate statistical analysis and insufficient sample sizes (Kilkenny et al., 2009; Prinz et al., 2011) putting the validity, replicability and translatability of results at stake.

The divergent results of the *Fmr1* KO phenotype raise questions about the validity as a preclinical model of neurodevelopmental disorders. In general, molecular studies quantifying the null expression of FMRP and its consequences on molecular alterations reach consensus. However, behavioural studies tend to show more discrepancies across laboratories and tasks. These discrepancies could suggest that the way in which the phenotype is assessed is not appropriate; for example, poorly sensitive tasks or poor experimental design, both of which are relevant for the internal and face validity of the model (Belzung and Lemoine, 2011; Campbell and Stanley, 1963). Additionally, it may be that the phenotype is not robust enough and therefore it only shows in some scenarios but can't be generalized to other study samples and/or scenarios, which questions the model's external validity (Campbell and Stanley, 1963; Richter, 2017). In order to objectively evaluate the phenotype of the *Fmr1* KO it is necessary to review the available literature and evaluate its methods.

Systematic review and meta-analysis are valuable tools to make a transparent (statistical) summary of research findings that yields an estimate of the validity of the overall findings. Although their use in preclinical science is relatively new, their value has been appraised by various disciplines such as medical sciences, psychology and education. By looking at the range of available published studies one can judge the external validity in addition to the possibility of assessing the risk of a publication bias. On the other hand, an indication of the internal validity based on a risk of bias assessment informs us about the methodological quality of the included studies overall (Sena et al., 2014). In this way, the systematic review and meta-analysis presented here will shed some light on the behavioural phenotype of the *Fmr1* KO line. Additionally, it will give insight into the experimental factors which affect the genotype expression and thereby potentially contribute to the variability in results presented in literature. In addition, an indication of the reporting quality in the field and a publication bias will be discussed further to properly ponder the results.

Given the large amount of available behavioural studies available in the literature, we decided to narrow down our systematic review and meta-analysis to the behavioural categories that are most relevant to evaluate the FXS/ASD-like phenotype. In the case of autism-like behaviours, we chose to focus on social behaviours, repetitive behaviours, anxiety, sensory gating, and sensory sensitivity as these are often reported in patients. In addition, learning, memory, and attention performance are relevant for the model given the intellectual disability component of FXS and thus, the *Fmr1* model (Harris et al., 2008). Locomotion was chosen based on its wide use given that most genetically modified models exhibit hyperactivity. Lastly, audiogenic seizures have high comorbidity with epilepsy as well as ASD and its increased excitation/inhibition (E/I) balance hypothesis.

METHODS

The review protocol was preregistered on PROSPERO (www.crd.york.ac.uk/prospero; CRD42020191070). The reporting in this systematic review adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist (Page et al., 2021; Supplementary file 12).

Search strategy

Two bibliographic databases were systematically searched for relevant studies: Pubmed and Web of Science. The search consisted of two components, one for fragile X syndrome and *Fmr1,* and one for rat and mouse (for full strategy see Supplementary file 1). If available, both controlled terms *(i.e.,* MeSH), and free text words were used. Bibliographic results were imported and de-duplicated using Rayyan software (Ouzzani et al., 2016). The final search was performed on 30-06-2021. In addition, the reference lists of all included studies were scanned for relevant studies that did not come up in the bibliographic search.

Eligibility screening

Studies were eligible for inclusion if they compared wildtype (WT) and knockout (KO) rodents for *Fmr1* in one or more behavioural tests relevant for our domains of interest (Supplementary file 2). The domains of interest included: locomotion, social behaviour (sociability, aggression, communication and social cognition), learning and cognition (conditioned learning, spatial learning, recognition learning and working memory), repetitive behaviour (low order repetitive behaviour and cognitive flexibility), anxiety, attention, sensory sensitivity (olfactory, somatosensory, auditory, visual and nociception) and sensitivity for audiogenic seizures. MA and RK independently screened all

identified records in two stages using Rayyan software. Disagreements were resolved by discussion.

The first stage concerned screening of the title and abstract of the articles. In this stage articles were excluded for the following reasons: (i) not an original primary study *(e.g.,* review, editorial or conference abstract), (ii) the used model was not a mouse or rat, (iii) the used model was not an *Fmr1* knockout (CGG-repeat knock-in models were not included), (iv) in vitro or ex vivo studies where behavioural assessment is impossible.
In the second stage, the full text of the remaining articles was screened. Articles were excluded for one or more of the reasons from stage one, plus the following additional reasons: (v) not a full KO *(e.g.,* selective or conditional KOs) or no use of WT control, (vi) no control condition for additional interventions present *(e.g.,* vehicle), (vii) behavioural tasks that did not fit with the behavioural domains of interest, (viii) no full text available.

Extraction of study characteristics

Extraction of study characteristics was performed by MA and RK, who both extracted characteristics for half of the studies. MA, a native speaker of the Spanish language, extracted the article written in Spanish. The following study characteristics were extracted: (i) study ID: first author, last author, year, journal, digital object identifier (DOI), article language; (ii) animal model characteristics: species, genetic background, sex, age, KO or KO2 (for mouse studies), being littermates; (iii) study design characteristics: housing conditions (group housed - mixed genotypes; group-housed - same genotypes; single housed), presence of additional interventions, test phase (light or dark), number of behavioural tasks; (iv) outcome measures: list of (relevant) behavioural tests used, list of test outcomes used.

Risk of bias assessment

Risk of bias assessment was performed to assess the methodological quality of the studies included in the meta-analysis. Due to the high number of papers included in the current study, and the high percentage of 'unclear risk of bias' scores expected because of poor reporting, the risk of bias analysis was performed on a random sample of 45 papers (18%). The SYRCLE risk of bias tool for animal studies (Hooijmans et al., 2014) was used. Item one and three from the tool, concerning randomization and blinding of treatment allocation, were only assessed for studies performing additional pharmacological interventions of which the vehicle groups were used in the current meta-analysis. Item seven (random outcome assessment) was scored as low risk of bias when data was scored using computerized automated scoring. Items were answered with a "Yes" for low risk of bias, "No" for high risk of bias and "Unclear" if it was not possible to assess the risk

of bias due to lack of information. Risk of bias assessment was independently performed by MA and RK, disagreements were resolved by discussion.

Extraction of outcome data

For every study data was extracted for each behavioural domain in which thebehavioural tests were performed. Mean, standard deviation (SD) or standard error (SEM) and the number of animals (N) were extracted for the WT and KO groups. For audiogenic seizures, the number of animals that did or did not experience seizures and the sample size was extracted for both WT and KO groups. If percentages of animals experiencing seizures were reported, the number of animals was calculated using the total sample size. Whenever possible, exact values were taken from text or tables. When those were not available, WebPlotDigitizer software (v3.8-4.4, Rohatgi, A., Pacifica, CA, USA, https://automeris.io/WebPlotDigitizer) was used to extract the numbers from figures. Although the initial protocol stated that authors would be contacted when using WebPlotDigitizer was not possible, we decided to refrain from this, due to the high number of studies and amount of data already included in the study. When it was unclear whether SD or SEM was reported, SEM was assumed, in order to be more conservative. Data extraction was performed by MA, RK and MAK. A random sample of studies (11, 4.3%) was extracted twice at the start to check referees' reliability. When ranges were reported for N, the highest value of the range was used to calculate the SD in case the study reported the SEM (), while the lowest value of the range was used as sample size in the actual meta-analysis (Ramsteijn et al., 2020). If a group of animals was used in comparison to multiple other groups *(e.g.,* WT females compared to both heterozygous and homozygous KO females), an adjusted sample size was used in the meta-analysis (sample size divided by the number of comparisons in which this group is used).

Before starting the extraction of outcome data, a categorization and prioritization of behavioural tests and outcomes was made by MA, RK and MAK and later discussed with MJHK and SB. All behavioural tests used within the included studies were allocated to one of the (sub-) domains of interest (Supplementary file 2). Within every (sub-)domain behavioural tests were ranked from most to least relevant, and for every test, outcome measures were ranked from most to least relevant. This ranking guided the data extraction, to assure that in the case of multiple reported outcomes or even multiple reported tests within the same behavioural domain, unique animals appeared only once in every domain. If a study performed experiments in multiple groups of animals *(e.g.,* males and females, different age groups or multiple additional interventions) we analysed these comparisons as if they were separate studies.

For social tasks, although social malfunctioning may be also expressed in male-female socio-sexual interactions and male-juvenile explorations across different ages, we decided to prioritize adult male-male interactions to characterise an adult phenotype independent of sexual and neurodevelopmental maturity. Furthermore, adult male-male interactions are the most frequently used for social interaction paradigms in studies on *Fmr1* KO mice (i.e., 45% of all reported social interaction, against 15% for male-juvenile, 10% for male-female and 8% for female-female). Ultrasonic vocalisation (USV) data was pooled over all call-types and frequencies. For cognitive tasks, including cognitive flexibility, the data from the last trial was always used to assess a stable outcome measure independent of the learning process. For recognition learning, the test with retention time closest to one hour was used. In the 5-choice serial reaction time task (5-CSRTT) the shortest stimulus duration was used. For acoustic startle and prepulse inhibition (PPI) responses, data was pooled over all tested startle intensities, prepulse intensities and inter-stimulus intervals, to have an unbiased assessment since studies show conflicting results across the range of startle and prepulse intensities (Baker et al., 2010; Braat et al., 2015; Ding et al., 2014; Hodges et al., 2019; Michalon et al., 2012; Naviaux et al., 2015; Zhang et al., 2014). For olfactory sensitivity tasks the data from the lowest concentration that was still detectable by WT animals was used. In the olfactory habituation-dishabituation task all first presentations, excluding water, were pooled. In the gap-crossing task, the data from gap-distances between five and six cm were pooled. For task assessing novelty recognition (social or object novelty recognition) data was only extracted when a ratio or index was reported, as the time spent interacting with the novel object/animal is only informative relative to the time spent interacting with the familiar object/animal. For locomotor activity, whenever available, only the first 30 minutes of exploration were extracted.

Meta-analysis

The meta-analysis was performed using comprehensive meta-analysis (CMA, v.3.3, Biostat Inc., Englewood, NJ, USA). For most outcome measures Hedge's G standardized mean differences (SMD) were used as the effect size measure. For the outcome measure audiogenic seizures, odds ratios were calculated. Because of anticipated heterogeneity, the effect sizes were pooled using a random effects model. Overall SMD were reported with 95% confidence intervals (CI). I^2 was used to assess statistical heterogeneity *(i.e.,* variation across studies due to heterogeneity rather than chance (Higgins and Thompson, 2002).

To further explore heterogeneity, subgroup analyses were performed. Subgroup analysis was only performed when there were at least 10 comparisons, from 5 unique studies for each subgroup. The original protocol listed only 5 comparisons from 3

unique studies, but based on the advice from the SYRCLE institute we increased these numbers to increase the power of the subgroup analyses. The effects of sex were explored by comparing studies only using male animals, with mixed sexes studies, using either females or both males and females, since there was not enough data to analyse females and male-female combined data separately. The effect of age was explored by grouping the age of experimental animals into juvenile (<6 weeks), adolescent (mice: 6-9 weeks; rats: 6-21 weeks) and adult (mice: >9 weeks, rats: > 21 weeks) (Adriani et al., 2004; Ghasemi, Asghar; Sajad, 2021; Semple et al., 2013; Sengupta, 2013). When an age-range was reported, the study was grouped into the age category the majority of the range belonged to. In cases were only the age at the start of testing was reported, this same age was taken when consecutive tests were performed. Genetic background effects were tested for C57/Bl6(J&N), FVB and FVBx129. Additionally, the effects of single vs group housing, being littermates or not and the KO vs KO2 mouse model were tested. Subgroup analyses were not performed on the other characteristics that were extracted because of a lack of data; this includes species (rat vs mouse) which was pre-specified in the initial protocol as a subgroup analysis factor. The number of behavioural tasks in a study was also prespecified as a subgrouping factor. However, during the extraction of the characteristics it turned out to be a complex outcome to extract due to various reasons, including the difficulty of defining when different phases of one task become separate tasks *(e.g.,* initial learning and reversal learning in the Morris water maze), missing information of whether or not tests were performed in different batches of animals, and uncertainty about how these would affect the meaning of this outcome. Thus, this characteristic was not taken into the meta-analyses as a subgrouping factor.
To test for differences between subgroups we calculated the confidence interval of the difference between the subgroups. Whenever three subgroups were compared, Bonferroni corrections were applied to correct for multiple comparisons. P-values lower than 0.05 were considered statistically significant.

Sensitivity analysis

A sensitivity analysis was performed to check if methodological or experimental differences reported between the studies could be skewing the main effect and should be considered separately. For this, the main effects of the meta-analysis when including all the studies were compared to the main effects when taking out those studies that reported different methodologies (*e.g.,* open field for more than 30 minutes) or did not explicitly report details that were assumed at the extraction phase. If the main effect remained unchanged after the removal of those studies, it was implied that those methodological differences were not dragging the meta-analysis main effect, therefore they could remain included. Experimental differences included, for example, assuming the error bars represented SEM when not specified, tests or stimuli with different time

lengths, whether data was pooled over stimuli or time, etc. See Supplementary file 7 for more details.

Publication bias assessment

Two different analyses were performed in parallel to assess whether meta-analyses showed significant asymmetry in the funnel plot and thus possibly suffered from publication bias, namely Egger's regression and Duval and Tweedie (Duval and Tweedie, 2000) trim and fill analysis (Stata Statistical Software, SE17, StataCorp LLC, College Station, TX). For both methods, the effect size estimate Hedges' G and sample-size based precision estimate $1/\sqrt{N}$ were used as it has been suggested that SE-based precision estimates cause distortion of SMD funnel plots (Wenstedt et al., 2021).

First, the Egger regression test was performed (Egger et al., 1997). This test is based on a simple linear regression and it can only identify small-study effects. In case of no publication bias, the regression line would cross the zero of the standard normal deviate *(i.e.,* precision estimate) in the y-axis.

Secondly, for the Duval and Tweedie, funnel plots were created where the effect sizes were plotted on the x-axis against $1/\sqrt{N}$ as a measure of precision on the y-axis (Zwetsloot et al., 2017). If there is no publication bias, studies are expected to spread equally across both sides of the overall effect size with larger deviations from the overall effect as the precision (*i.e.,* sample size) of the study decreases.

All data is publicly available in supplementary files via the OSF repository (https://osf.io/d2cbx/), this includes all the extracted data (Supplementary file 8), the statistical results of the meta-analysis (Supplementary file 9), the statistical results of the subgroup analyses (Supplementary file 10) and the statistical results of the publication bias analysis (Supplementary file 11). Additionally, on this repository the methods and results of the behavioural experiments we performed in our own lab can also be found.

RESULTS

Search results

In total, 5065 records were retrieved through database screening. After duplicate removal 3414 unique records were scanned for eligibility. Via title and abstract screening, 374 records were selected for full-text assessment. Of those, 265 articles were found to be eligible for inclusion in the systematic review and risk of bias assessment, together with one study that was found by scanning the reference lists of included articles. From the

266 studies of the systematic review, 15 studies were excluded from the meta-analyses because they did not contain the right data (Fig. 1). Thus, 251 studies were included in the meta-analysis, as the minimum of five independent studies was reached for every behavioural domain. The digital object identifiers (DOIs) of all included studies can be found in the characteristics table (Supplementary file 3).

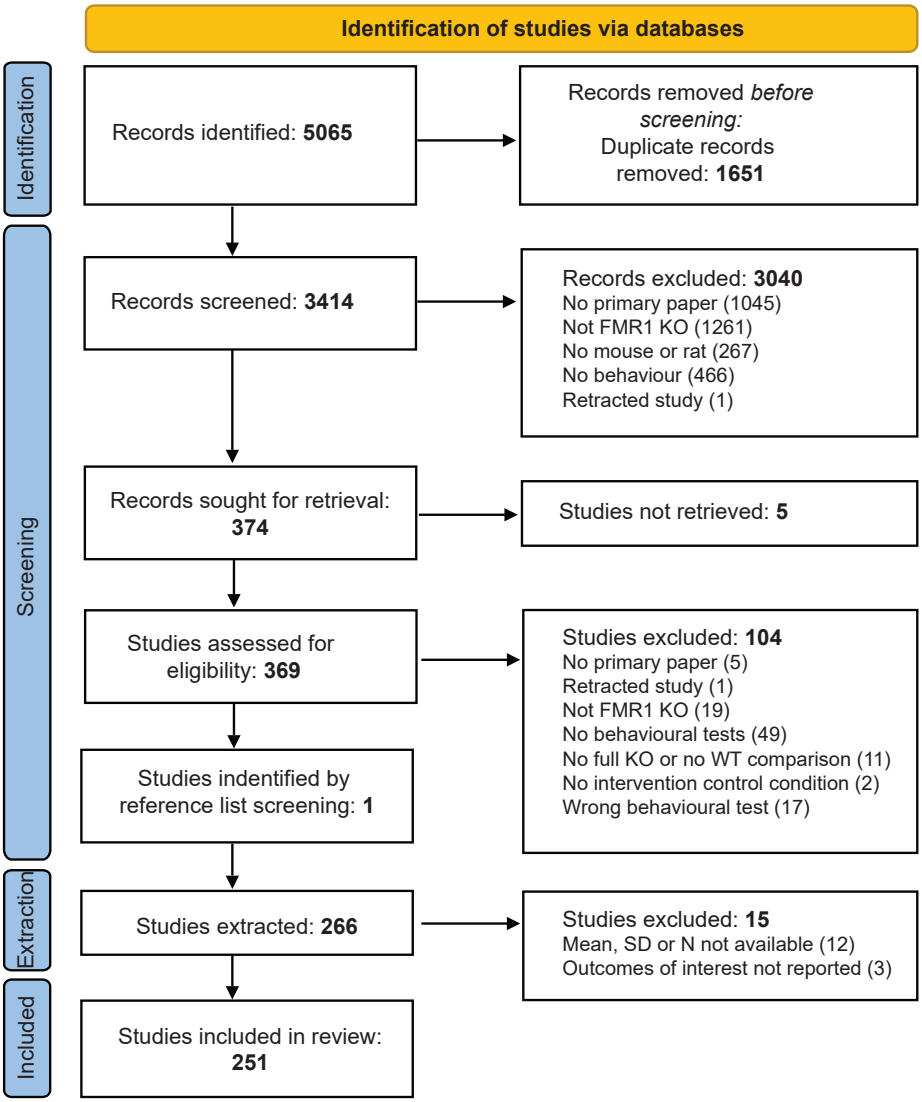


Fig 1. Study flowchart. All behavioural categories reached the minimum number of studies needed for meta-analysis, therefore all studies included in the systematic review were also included in the meta-analysis.

Study Characteristics

The characteristics of all included studies can be found in supplementary file 3. From the 266 included studies 252 used mice, 10 used rats and one study used both rats and mice. Of the studies performed in mice, the C57BL/6 background was the most frequently used background (151), but also FVB (80) and FVBx129 (9) were used frequently. Twenty-five studies used other backgrounds and six studies did not report the genetic background of the mice. In rat studies, both Sprague-Dawley (7) and Long Evans (4) backgrounds were used. Data was reported either specifically for males (215) and females (20), or for the two sexes combined (23). The sex of the animals was not specified in 22 studies. The majority of studies used adult animals (173), followed by juvenile (63) and adolescent (42) animals. In 23 studies, the age of the animals was not reported. From the studies using mice, 179 used the first-generation KO, 19 used the second-generation KO (KO2) and 65 did not specify which model was used. Most studies tested KO and WT animals as littermates (160), but in 33 studies control animals were not littermates and in 74 studies it was not reported. In most studies animals were group-housed (179), 26 of which used housing with mixed genotypes, 11 with the same genotype and for 142 studies it was unknown how the groups were composed. In 18 studies animals were individually housed during experiments and 123 studies did not report on housing conditions. The majority of studies did not specify whether behavioural tests were performed during the light phase or dark phase. From the studies that did report the testing phase, 116 performed tests during the light phase, 11 during the dark phase and three performed 24h recordings, thus including both light phases.

None of the experimental or methodological differences tested for in the sensitivity analysis affected the main effect in any of the meta-analyses.

Study Quality

A risk of bias assessment was performed according to the SYRCLE's RoB tool (Hooijmans et al., 2014) in a random subset of the included articles (Supplementary file 4). Overall, the risk of bias in these articles was unclear (Fig. 2). Blinded execution was reported in 49% of the articles and 58% of the studies assessed the outcomes blinded for genotype, while one study reported to not be blinded (2%). Except for one study stating that outcome assessment was not performed in a randomized order, none of the studies mentioned randomization of outcome assessment. Outcome data was incomplete in four studies (9%) and it was unclear whether data was complete in 78% of assessed studies. Five studies (11%) did not report on all the outcomes presented in the methods section.

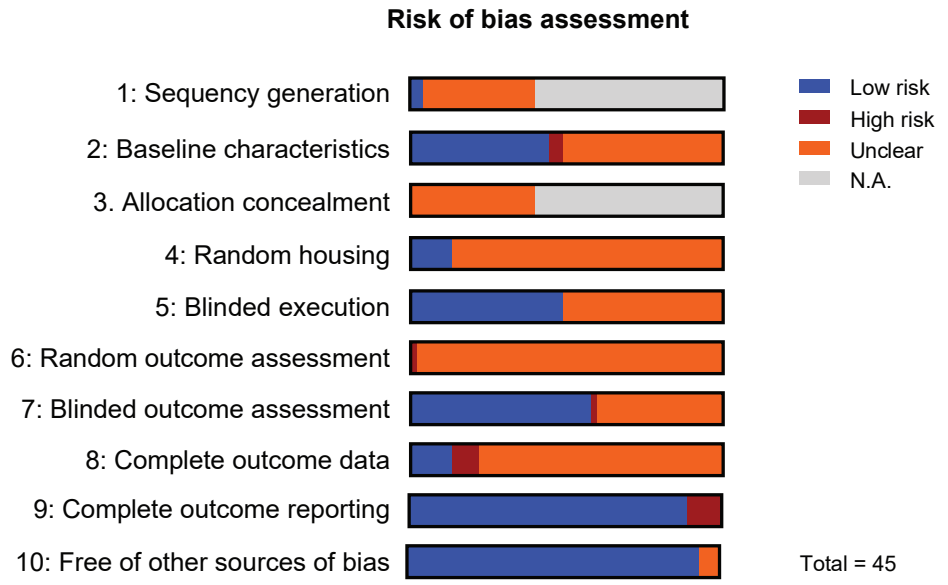


Fig 2. Risk of bias assessment outcomes. Risk of bias assessment was performed with SYRCLE's risk of bias assessment tool. Item 1 and 3 were only used in studies with an intervention, in which was assessed if animals were randomly assigned to the control condition. Risk of bias analysis was performed on a random sample of 45 studies by two independent assessors.

Locomotion

The meta-analysis comprised 176 comparisons out of 125 independent studies. A total of 2331 WT and 2299 KO animals were included in the analysis. The most frequently used behavioural test to assess locomotion was the open field test (145), followed by the three-chamber test (6), Novel Object Recognition Test (NORT) training phase (3), actimetry cages (3), Morris water maze (2), home cage activity (1), active place avoidance (1), elevated plus maze (1), plus-shaped water maze (1) and visual cliff (1).

Ninety-two comparisons had a point estimate significantly larger than zero, four comparisons had a point estimate significantly smaller than zero and 80 comparisons did not significantly deviate from zero. Overall analysis showed that *Fmr1* KO animals have a significant increase in locomotor activity compared to WT controls (SMD 1.046 [0.878, 1.214], $P < 0.001$, $I^2 = 85.4$, Fig. 3, Supplementary file 5). The heterogeneity was considerable and remained unchanged after the subgroups analyses.

The genotype effect did not differ between genetic backgrounds (B6 vs FVB: t(132) = 2.08, $P = 0.12$; B6 vs FVBx129: t(121) = 0.64, $P = 1.56$; FVB vs FVBx129: t(51) = 0.99, $P = 0.99$), sexes (t(161) = 0.90, $P = 0.37$), age groups (Juvenile vs Adolescent: t(54) = 1.35; $P = 0.55$); Juvenile vs Adult: t(121) = 0.67; $P = 1.52$; Adolescent vs Adult: t(145) = 1.04,

P = 0.90), littermates and non-littermates (t(124) = 0.58, *P* = 0.56), the first and second generation KO (t(128) = 1.09, *P* = 0.28) nor single and group housed animals (t(114) = 1.77, *P* = 0.080).

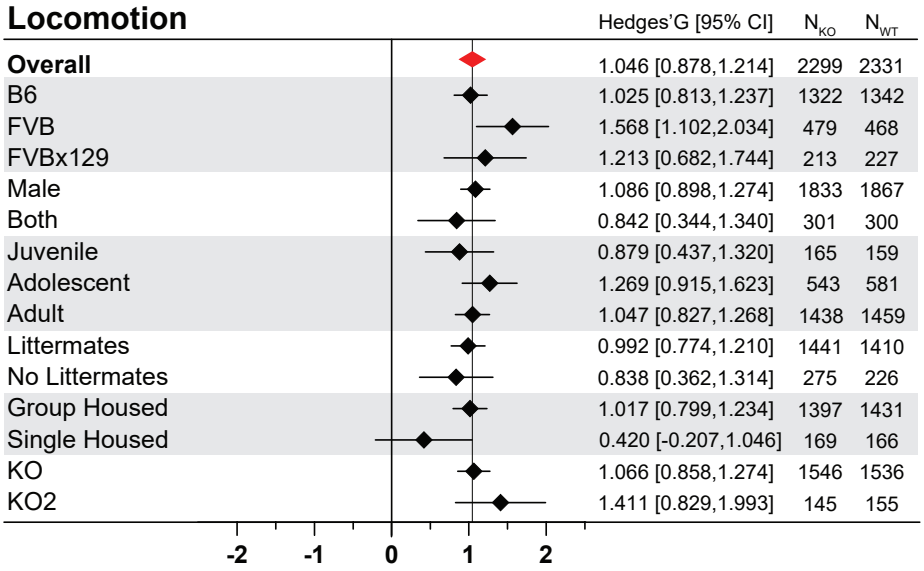


Fig 3. The effect of Fmr1 KO on locomotor activity. Subgroup analyses were performed for genetic background, sex, age, littermates, housing condition and KO line. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N_{KO} and N_{WT}) included in each of the comparisons.

Cognition

Conditioned Learning

The meta-analysis comprised 134 comparisons out of 83 independent studies. A total of 1752 WT and 1791 KO animals were included in the analysis. The most frequently used behavioural test to assess conditioned learning was fear conditioning (70), followed by passive avoidance (35), discrimination learning (13), active avoidance (6), operant conditioning (6), conditioned place preference (3) and conditioned taste aversion (1).

Sixty-four of the comparisons had a point estimate significantly smaller than zero, three comparisons had a point estimate significantly larger than zero and 67 comparisons did not significantly deviate from zero. Overall, *Fmr1* KO animals show a significant decrease in conditioned learning compared to WT controls (SMD -0.862 [-1.023, -0.702], *P* < 0.001, I^2 = 80.3 Fig. 4, Supplementary file 5), however there was high heterogeneity which did not decrease with subgroup analysis.

The difference between KO and WT animals seemed larger in the FVBx129 and FVB backgrounds compared to the B6 background, although not significantly (B6 vs FVBx129: t(84) = 1.92, *P* = 0.18; B6 vs FVB: t(99) = 2.07, *P* = 0.12; FVB vs FVBx129: t(37) = 0.03, *P* = 2.91). The genotype effect was larger in animals that were not littermates, compared to animals that were littermates (t(93) = 2.56, *P* = 0.012). The genotype effect did not differ between sexes (t(121) = 0.98, *P* = 0.33), nor age groups (Juvenile vs Adolescent: t(28) = 0.28, *P* = 2.34; Juvenile vs Adult: t(107) = 1.01, *P* = 0.94; Adolescent vs Adult: t(109) = 1.24, *P* = 0.65).

Spatial Cognition

The meta-analysis comprised 69 comparisons out of 42 independent studies. A total of 725 WT and 752 KO animals were included in the analysis. The most frequently used behavioural test to assess spatial memory was object location memory (24), followed by the Morris water maze (17), categorical spatial processing task (6), plus-shaped water maze (5), radial maze (5), non-match to place learning (4), y-maze (3), Barnes maze (2), E-maze (1), Hebb-William maze (1) and metric change in the NORT (1).

Thirty-three out of these comparisons had a point estimate significantly smaller than zero, two had a point estimate significantly larger than zero and 34 comparisons did not deviate from zero. *Fmr1* KO animals show a robust and significant impairment in spatial cognition compared to WT controls (SMD -0.956 [-1.197, -0.715], *P* < 0.001, I^2 = 79.1, Fig. 4, Supplementary file 5). Heterogeneity did not reduce with subgroup analysis.

The genotype effect did not differ between animals that were littermates or no littermates (t(58) = 0.35, *P* = 0.73).

Recognition Learning

The meta-analysis comprised 53 comparisons out of 34 independent studies. A total number of 604 WT and 519 KO animals were included in the analysis. All studies used the NORT, two of which used the temporal order version of the NORT.

Forty out of these comparisons had a point estimate significantly smaller than zero and 13 comparisons had a point estimate not significantly different from zero. *Fmr1* KO animals show a robust and significant impairment in recognition memory compared to WT controls (SMD -1.696 [-2.025, -1.367], *P* < 0.001, I^2 = 82.5, Fig. 4, Supplementary file 5). Heterogeneity did not reduce with subgroup analysis.

Working Memory

The meta-analysis comprised 15 comparisons out of 13 independent studies. A total of 183 WT and 187 KO animals were included in the analysis. The most frequently used behavioural task to assess working memory was spontaneous alternations in the Y-maze (6) and the T-maze (6), followed by working memory errors in radial arm maze learning (2), delayed non-match to place learning (1) and serial reversals in the Morris water maze (1).

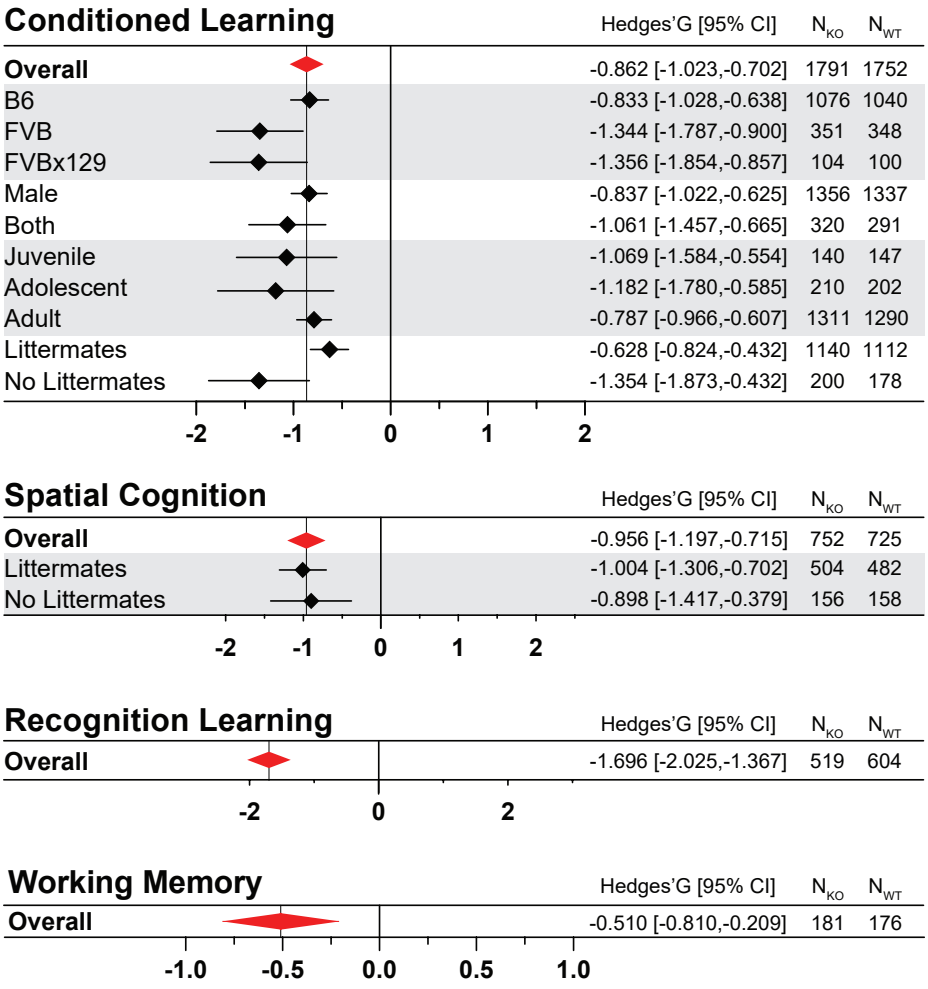


Fig 4. The effect of Fmr1 KO on cognition. Meta-analyses were performed in the category of conditioned learning, spatial cognition, recognition learning and working memory. Subgroup analyses were performed for genetic background, sex, age, and/or littermates. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N_{KO} and N_{WT}) included in each of the comparisons. X-axis limits differ per meta-analysis to optimize subgroup error band visualization.

Three out of these comparisons had a point estimate significantly smaller than zero, while 12 comparisons did not deviate from zero. *Fmr1* KO animals show a significant impairment in working memory compared to WT controls (SMD -0.510 [-0.810, -0.209], $P = 0.001$, $I^2 = 49.7\%$ Fig. 4, Supplementary file 5).

Repetitive behaviour

Low order repetitive behaviour

The meta-analysis comprised 87 comparisons out of 53 independent studies. A total of 1063 WT and 1098 KO animals were included in the analysis. The most frequently used behavioural test to assess low order repetitive behaviour was the marble burying test (42), followed by spontaneous behaviour in the open field (33), fear conditioning (3), three-chamber (2), y-maze (1) or elevated plus maze (1), block chew test (2) and the nose-poke assay (2).

Thirty-six out of these comparisons had a point estimate significantly larger than zero, nine comparisons had a point estimate significantly smaller than zero and 42 comparisons did not deviate from zero. *Fmr1* KO animals show a significant increase in low order repetitive behaviours compared to WT controls (SMD 0.572 [0.356, 0.789], $P < 0.001$, $I^2 = 82.4$, Fig. 5, Supplementary file 5), however there was considerable heterogeneity which did not decrease with subgroup analysis.

The genotype effect did not differ between genetic backgrounds (B6 vs FVB: $t(69) = 0.25$, $P = 0.81$), sex ($t(82) = 0.78$, $P = 0.44$), age groups (Juvenile vs Adolescent: $t(23) = 0.51$, $P = 1.84$, Juvenile vs Adult: $t(70) = 0.81$, $P = 1.27$; Adolescent vs Adult: $t(73) = 1.56$, $P = 0.34$), nor sexes ($t(82) = 0.78$, $P = 0.44$).

Cognitive Flexibility

The meta-analysis comprised 30 comparisons out of 23 independent studies. A total of 352 WT and 361 KO animals were included in the analysis. The most frequently used behavioural test to assess cognitive flexibility was reversal in the Morris water maze (10), followed by discrimination learning reversal (4), plus-shaped water maze reversal (3), y-maze reversal (3), passive avoidance extinction (3), active avoidance extinction (3), operant conditioning extinction (1), fear conditioning extinction (1), E-maze reversal (1) and 5-CSRRT reversal (1).

Eight out of these comparisons had a point estimate significantly smaller than zero, four had a point estimate significantly larger than zero and 18 comparisons did not devi-

ate from zero. *Fmr1* KO animals do not show a cognitive flexibility deficit (SMD -0.179 [-0.483, 0.125], $P < 0.249$, $I^2 = 75.0$, Fig. 5, Supplementary file 5).

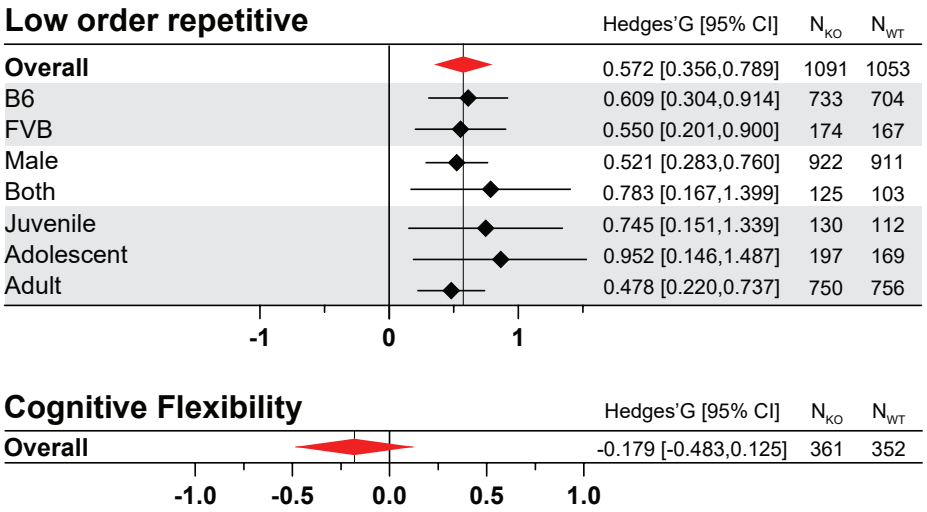


Fig 5. The effect of Fmr1 KO on repetitive and restricted behaviour. Meta-analyses were performed in the category of low order repetitive behaviour and cognitive flexibility. Subgroup analyses were performed for genetic background, sex and age. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N_{KO} and N_{WT}) included in each of the comparisons. X-axis limits differ per meta-analysis to optimize subgroup error band visualization.

Social behaviour

Sociability

The meta-analysis comprised 107 comparisons out of 69 independent studies. A total of 1424 KO and 1399 WT animals were included in the analysis. The most frequently used behavioural task to assess sociability was the three-chamber test (67), followed by the direct social interaction test (23), partition test (11), tube co-occupancy test (2), resident-intruder test (2), Eco-HAB (1) and the social conditioned place preference test (1).

Thirty-six out of these comparisons had a point estimate significantly smaller than zero, 10 comparisons had a point estimate significantly larger than zero and 61 studies did not deviate from zero. *Fmr1* KO animals show a significant decrease in sociability compared to WT controls (SMD -0.368 [-0.546, -0.189], $P < 0.001$, $I^2 = 81.1$, Fig. 6, Supplementary file 5), however there was a high degree of heterogeneity which did not reduce in subgroup analysis.

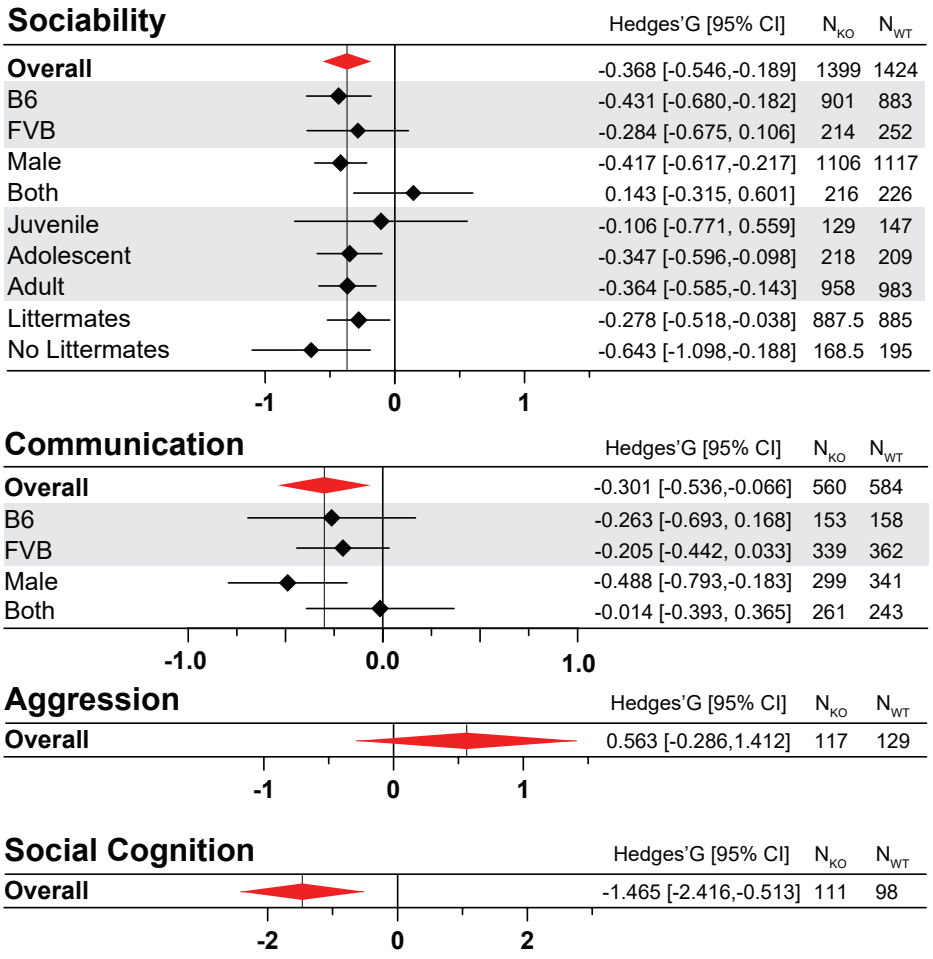


Fig 6. The effect of Fmr1 KO on social behaviour. Meta-analyses were performed in the category of sociability, communication, aggression and social cognition. Subgroup analyses were performed for genetic background, sex, age and littermates. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N_{KO} and N_{WT}) included in each of the comparisons. X-axis limits differ per meta-analysis to optimize subgroup error band visualization.

The genotype effect was significantly larger in studies using only male animals compared to studies using both sexes, in which the effect was also not significantly different from zero (SMD 0.143 [-0.315 0.601], t(99) = 2.19, *P* = 0.030). The genotype effect did not differ between genetic backgrounds (B6 vs FVB: t(86) = 0.62, *P* = 0.53), age groups (Juvenile vs Adolescent: t(25) = 0.8673, *P* = 1.54; Juvenile vs Adult: t(83) = 0.72, *P* = 1.42; Adolescent vs Adult: t(84) = 0.10, *P* = 2.76), littermates and non-littermates (t(78) = 1.39, *P* = 0.17).

Communication

The meta-analysis comprised 35 comparisons out of 21 independent studies. A total of 584 WT and 560 KO animals were included in the analysis. Most USVs were isolation-induced (22) or socially-induced (10). USVs were also recorded in the resident-intruder test (2) and the open field (1).

Ten out of these comparisons had a point estimate significantly smaller than zero, four comparisons had a point estimate larger than zero and 21 comparisons did not deviate from zero. *Fmr1* KO animals show a significant communication deficit (SMD -0.301 [-0.536, -0.066], $P = 0.127$, $I^2 = 72.1$, Fig. 6, Supplementary file 5). The overall heterogeneity did not reduce in subgroup analysis.

The genotype effect was only present in studies using only males (-0.488 [-0.793, -0.183]), and not in studies using both sexes (-0.014 [-0.393, 0.365]), although the difference between the sexes was not significant ($t(33) = 1.91$, $P = 0.065$). The genotype effect did not differ between genetic backgrounds ($t(27) = 0.23$, $P = 0.82$).

Aggression

The meta-analysis comprised 10 comparisons out of six independent studies. A total of 129 WT and 117 KO animals were included in the analysis. The most frequently used test to assess aggressive behaviour was the direct social interaction task (6) followed by the tube test (3) and the dominance hierarchies (1).

Five out of these comparisons had a point estimate significantly larger than zero, two had a point estimate significantly larger than zero and three comparisons did not significantly deviate from zero. *Fmr1* KO animals did not show enhanced aggression (SMD 0.563 [-0.286, 1.412], $P = 0.194$, $I^2 = 89.6$, Fig. 6, Supplementary file 5).

Social Cognition

The meta-analysis comprised 10 comparisons out of seven independent studies. A total of 98 WT and 111 KO animals were included in the analysis. All assessments of social cognition were performed in the three-chamber test. Eight out of these comparisons had a point estimate significantly smaller than zero and two comparisons did not deviate from zero. *Fmr1* KO animals showed a consistent significant reduction in social cognition (SMD -1.465 [-2.416, -0.513], $P = 0.003$, $I^2 = 87.9$, Fig. 6, Supplementary file 5).

Anxiety

The meta-analysis comprised 136 comparisons out of 96 independent studies. A total of 1838 WT and 1882 KO animals were included in the analysis. The most frequently used

behavioural test to assess anxiety was the open field (58), followed by the elevated plus maze (36), the light-dark test (32), the elevated zero maze (6), the successive alleys maze (2), the mirrored chamber (1) and the platform test (1).

Fifty-two out of these comparisons had a point estimate significantly smaller than zero, seven comparisons had a point estimate significantly larger than zero and 77 studies did not deviate from zero. *Fmr1* KO animals show a significant decrease in anxiety compared to WT controls (SMD -0.555 [-0.692, -0.419], $P < 0.001$, $I^2 = 75.1$, Fig. 7, Supplementary file 5), The overall heterogeneity did not decrease with subgroup analysis.

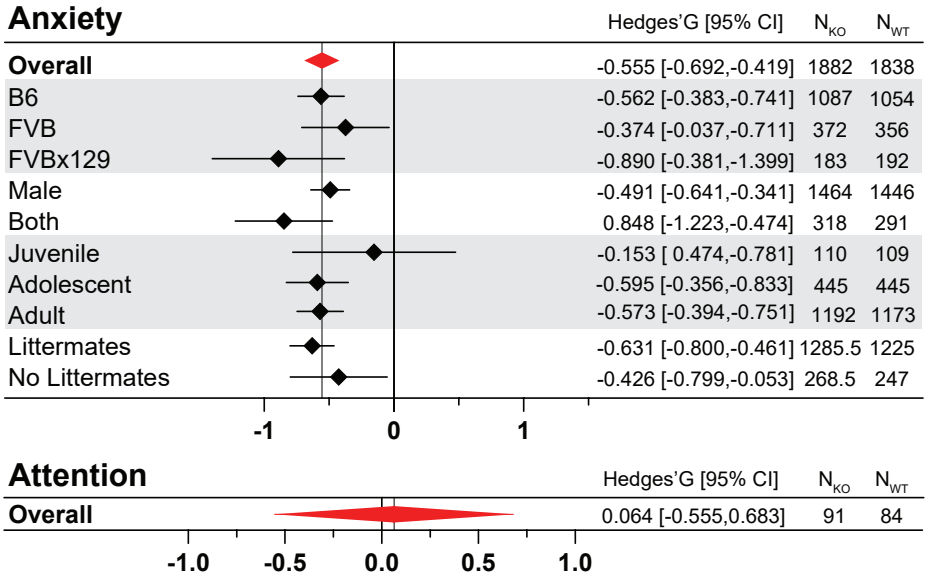


Fig 7. The effect of Fmr1 KO on anxiety and attention. Subgroup analyses were performed for genetic background, sex, age and littermates. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N_{KO} and N_{WT}) included in each of the comparisons. X-axis limits differ per meta-analysis to optimize subgroup error band visualization.

The genotype difference was smaller in juvenile compared to adolescent and adult animals, although this difference was not significant (Juvenile vs Adolescent: t(39) = 1.29, $P = 0.61$; Juvenile vs Adult: t(96) = 1.26, $P = 0.63$; Adolescent vs Adult: t(115) = 0.15, $P = 2.65$). Similarly, although the effect was larger in studies using both sexes, this difference did not reach statistical significance (t(127) = 1.73, p = 0.086). The difference between KO and WT animals was not affected by genetic background (B6 vs FVB: t(98) = 0.97, $P = 1.01$; B6 vs FVBx129: t(92) = 1.12, $P = 0.71$; FVB vs FVBx129: t(40) = 1.66, $P = 0.32$), nor littermates and non-littermates (t(106) = 0.98, $P = 0.33$).

Attention

The meta-analysis comprised seven comparisons out of five independent studies. A total of 84 WT and 91 KO animals were included in the analysis. All assessments of attention were performed in the 5-choice serial reaction time task. One of the comparisons had a point estimate significantly smaller than zero, one had a point estimate significantly larger than zero and five comparisons did not deviate significantly from zero. *Fmr1* KO animals did not show an attention deficit (SMD 0.064 [-0.555, 0.683], $P = 0.839$, $I^2 = 75.5$, Fig. 7, Supplementary file 5).

Startle and prepulse inhibition

Acoustic Startle

The meta-analysis comprised 56 comparisons out of 40 independent studies. A total of 883 WT and 866 KO animals were included in the analysis. Nineteen out of these comparisons had a point estimate significantly smaller than zero, seven comparisons had a point estimate significantly larger than zero and 30 comparisons did not deviate from zero. *Fmr1* KO animals show a significantly decreased acoustic startle compared to WT controls (SMD -0.335 [-0.591, -0.079], $P = 0.010$, $I^2 = 84.9$, Fig. 8, Supplementary file 5).

The startle deficit was present in mice with a FVB background (-0.838 [-1.242, -0.434]), but not in mice with a B6 background (-0.045 [-0.471 0.381], $t(36) = 2.65$, $P = 0.012$). The overall heterogeneity did not reduce with subgroup analysis.

Prepulse Inhibition

The meta-analysis comprised 46 comparisons out of 30 independent studies. A total of 613 WT and 598 KO animals were included in the analysis. Twenty out of these comparisons had a point estimate significantly larger than zero and 26 comparisons did not deviate from zero. *Fmr1* KO animals show a significantly increased prepulse inhibition compared to WT controls (SMD 0.601 [0.403, 0.799], $P < 0.001$, $I^2 = 65.2$, Fig. 8, Supplementary file 5). Although effects in the opposite direction were not found, the heterogeneity was still considerable. The genotype effect did not differ between genetic backgrounds ($t(29) = 0.15$, $P = 0.89$).

Sensory sensitivity

The meta-analysis comprised 41 comparisons out of 26 independent studies. A total of 525 WT and 484 KO animals were included in the analysis. The most frequently used behavioural test to assess sensory sensitivity was the hot plate (16), followed by chemically-induced pain (5), odour habituation-dishabituation test (4), odour discrimination (3), buried food test (2), von Frey test (2), gap crossing task (2), olfactory sensitivity test (2), whisker-dependent tex-

ture discrimination (2), visual cliff test (1), texture NORT (1) and shock sensitivity (1). Eight out of these comparisons had a point estimate significantly smaller than zero, one comparison had a point estimate significantly larger than zero and 32 comparisons did not deviate from zero. *Fmr1* KO animals show a significantly decreased sensory sensitivity compared to WT controls (SMD -0.412 [-0.586, -0.239], $p < 0.001$, $I^2 = 46.7$, Fig. 8, Supplementary file 5). The overall heterogeneity did not decrease with subgroup analysis.

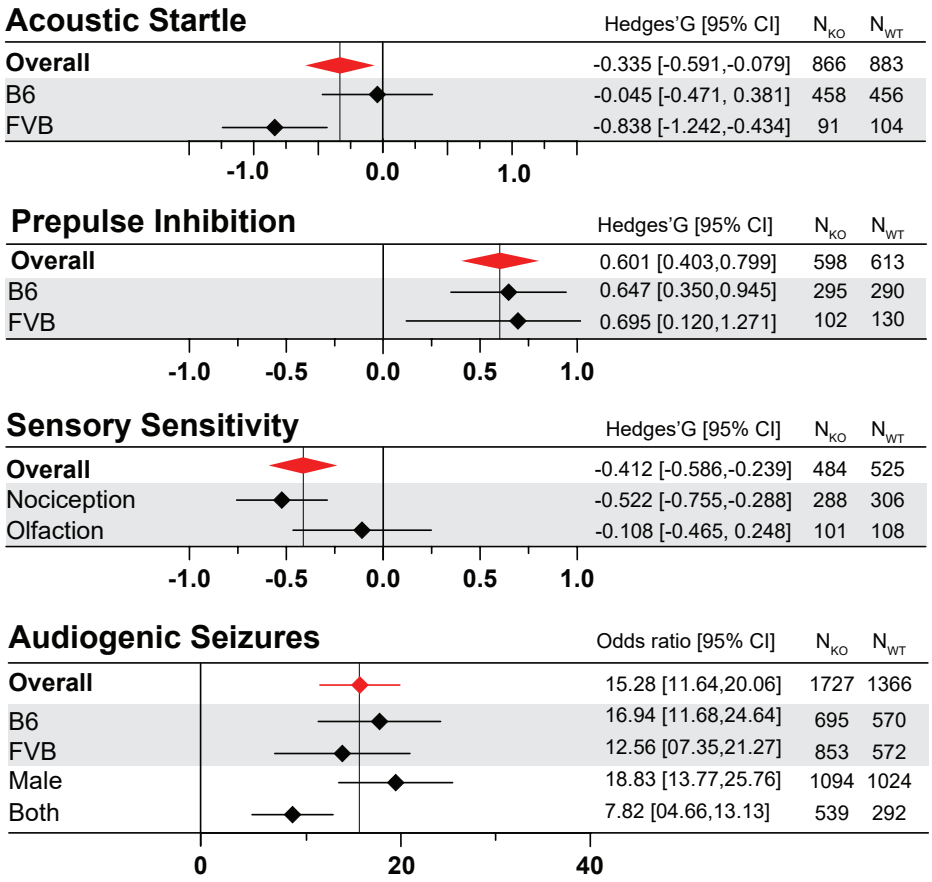


Fig 8. The effect of Fmr1 KO on sensory processing. Meta-analyses were performed in the category of acoustic startle, prepulse inhibition, sensory sensitivity and audiogenic seizures. Subgroup analyses were performed for genetic background, sex and sensory modality. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N_{KO} and N_{WT}) included in each of the comparisons. X-axis limits differ per meta-analysis to optimize subgroup error band visualization. Note that the results of the audiogenic seizures (bottom panel) are expressed in odds ratio.

The sensory sensitivity deficit seemed to be stronger for nociception compared to olfaction, though not significantly ($t(31) = 1.96$, $p = 0.059$).

Audiogenic seizures

The meta-analysis comprised 98 comparisons out of 40 independent studies. A total of 1376 WT and 1737 KO animals were included in the analysis. Sixty-three out of these comparisons had a point estimate significantly larger than zero, 35 comparisons did not deviate from zero. *Fmr1* KO animals show a robust and significant increased sensitivity for audiogenic seizures (Odds ratio 15.280 [11.643, 20.055], $P < 0.001$, $I^2 = 11.7$, Fig. 8, Supplementary file 5). The overall heterogeneity did not decrease with subgroup analysis.

The genotype effect was larger in studies using only male animals compared to studies using both sexes ($t(92) = 2.47$, $P = 0.015$). The genetic background did not affect seizure sensitivity (B6 vs FVB: $t(83) = 0.73$, $P = 0.46$.

Publication bias

Publication bias was assessed through funnel plot's asymmetry according to Egger's regression test for small-study effects supplemented with Duval and Tweedie trim and fill analysis. The meta-analysis for audiogenic seizures was performed using odds ratio and was therefore assessed for publication bias only with a trim and fill analysis.

Inspection of the funnel plots did not reveal asymmetry with either the Egger test nor Duval and Tweedie test for the anxiety, aggression, conditioned learning, cognitive flexibility, communication, locomotion, PPI, sensory sensitivity, sociability, and social cognition (Supplementary file 11).

Egger's regression test indicated bias for three behavioural categories: acoustic startle ($P = 0.005$), recognition learning ($P = 0.012$), and spatial cognition ($P = 0.040$) (Supplementary file 11).

The Duval and Tweedie test showed funnel plot asymmetry for two categories, namely acoustic startle and low order repetitive behaviour. For the acoustic startle response, studies showing increased startle response by the KO animals compared to WT were underrepresented. This resulted in 15 imputed studies and an adjusted effect size estimate of 0.045 [-0.249, 0.339] (Supplementary file 6). In the low order repetitive behaviour, studies showing decreased repetitive results were underrepresented leading to imputing 27 extra studies resulting in an adjusted effect size to -0.031 [-0.370, 0.307] (Supplementary file 6). For these two categories, the direction of the effect size changed after adjusting for the trim and fill analysis. These results should therefore be cautiously interpreted as marginal effects could be inflated by publication bias.

Additionally, the trim and fill analysis using odds ratios for the audiogenic seizures also showed funnel plot asymmetry. Twenty-four extra studies were added and gave an adjusted effect size of 3.917 [3.213, 4.774] (Supplementary file 6). The effect size direction remained the same after adjustment.

The discrepancies shown by these two publication bias analysis methods could be explained by the different methodologies they use. However, both methods indicated a significant overestimation of the genotypic effect in the acoustic startle response due to publication bias.

DISCUSSION

In the current systematic review and meta-analysis, we aimed to shed light on the behavioural profile of the *Fmr1* KO model, how it matches the clinical manifestation of FXS and which experimental factors might explain the heterogeneity of results seen in literature. We were able to include a large body of literature, which allowed us to perform meta-analyses in all relevant behavioural categories; however, in preclinical meta-analyses there is a trade-off between power and heterogeneity, which makes correct interpretation of the overall effect more complex. Irrespective of the overall effects found, this meta-analysis underscores the large inconsistencies between studies with effects being replicated in less than 50% of the independent comparisons in 10 out of 14 categories. This heterogeneity could represent true between-study variation in study design and experimental conditions (*i.e.,* phenotypic flexibility due to environmental diversity), but this was hard to assess due to the poor reporting of experimental factors. Additionally, low sample sizes and suboptimal research practices are likely to contribute to the low replicability of the phenotypes as studies with higher effect sizes showed more consistent results. Both incomplete reporting of experimental methods and conditions, and underpowered studies are common problems in behavioural preclinical neuroscience research (Sena et al., 2014). This meta-analysis stresses the need for improvements, not only regarding the *Fmr1* KO, but for animal research in the preclinical field as a whole.
With the estimated overall effect sizes that resulted from the meta-analysis, we were able to look at achieved power and required sample sizes in the various behavioural categories. Based on the estimated overall effect size, required sample sizes for sociability and anxiety would be more than 100 animals per genotype in order to reach a statistical power of 0.8. Additionally, when calculating achieved power based on the estimated overall effect size and the average sample size, only recognition learning and social cognition reach a power of at least 0.8. However, computed achieved power does not match perfectly with the percentage of studies replicating certain effects, indicating

that power is probably not the only factor causing the inconsistency of results. Also, these post-hoc calculations must be interpreted carefully, since due to the diverse experimental designs and research practices across studies, they cannot be translated to specific experimental settings. Nevertheless, insufficient power and sample sizes should be addressed as contributing factors to the low replicability of results.

It has also been reported that the rigorous standardization of animal experiments can lead to behavioural findings that can be replicated only under the exact same environmental and experimental conditions, which limits the interpretation and replicability of results (Richter, 2017; Voelkl and Würbel, 2016; Wurbel, 2000). FXS, like most neuropsychiatric disorders, is a complex disease where patients show high variability of phenotypes in terms of their symptoms and their severity (Ciaccio et al., 2017; Jacquemont et al., 2014). Likewise, animal models have shown phenotypic flexibility and so, the inconsistency of results between preclinical studies may be partly explained by the restricted generalizability and accuracy of results from study to study. Incorporating controlled biological variation into animal experiments could increase the external validity of findings (Voelkl et al., 2020). In addition, multicentre studies (Inthout et al., 2016) or multi-batch studies (Karp et al., 2020) are recommended in order to increase the robustness of studies assessing behavioural phenotype of animal models as these experimental designs have proven to render more representative study samples which allows more generalizable results. This could contribute to higher consistency across findings and thus more conclusive results.

Despite the large heterogeneity, we found significant overall effects matching the direction of the clinical profile in the majority of behavioural categories (Table 1). However, no effects were found on cognitive flexibility, attention and aggression although patients show flexibility and attention deficits, and enhanced aggression (Table 1, in bold). Nevertheless, these meta-analyses which did not show effects had a relatively low number of studies and total number of animals, and sometimes large confidence intervals, so results should be interpreted carefully. On the other hand, the reduced anxiety and acoustic startle, and enhanced PPI found in the KO animals are even opposite to the symptoms seen in patients. Strikingly, in patients the prevalence of problems with attention (74-84%), aggression (90%) and anxiety (58-86%) are higher than the prevalence of ASD (30-50%) and epilepsy (10-20%) of which the social, repetitive and seizure phenotypes were captured in the KO animals (Ciaccio et al., 2017).

There are multiple possible explanations for the phenotype mismatch between the meta-analysis and the clinical population considering anxiety, startle and PPI. True species-specific differences in the mechanisms and thus the way the disorder presents

Table 1. Comparison of the meta-analysis findings to the clinical phenotype.

Behavioural category	Meta-Analysis	Clinical Phenotype
Locomotion	↑	↑[1]
Conditioned learning	↓	↓[1,2]
Spatial cognition	↓	↓[1,3]
Recognition learning	↓	↓[1,4]
Working memory	↓	↓[5]
Low order repetitive	↑	↑[6]
Cognitive flexibility	=	↓[5]
Sociability	↓	↓[6]
Communication	↓	↓[6]
Aggression	=	↑[1]
Social cognition	↓	
Anxiety	↓	↑[1]
Attention	=	↓[1,5]
Acoustic startle	↓	↑[7-10]
PPI	↑	↓[7-10]
Sensory sensitivity	↓	↓↑[11]
Audiogenic seizures	↑	↑[1]

Categories in which the findings of the meta-analysis do not match the clinical phenotype are printed in bold text. [1](Ciaccio et al., 2017), [2](Reeb-Sutherland and Fox, 2015), [3](MacLeod et al., 2010), [4](Kogan et al., 2009), [5](Schmitt et al., 2019), [6](Niu et al., 2017), [7](Berry-Kravis et al., 2009), [8](Frankland et al., 2004), [9](Hessl et al., 2009), [10](Yuhas et al., 2011), [11](Baranek et al., 2009).

itself in rodents and humans may exist. Discrepancies in anxiety findings might also result from the challenging assessment and interpretation of this complex behaviour in rodents. For example, some drugs known to be anxiolytic in humans are ineffective or even anxiogenic in the open field test, the most frequently used anxiety test in this meta-analysis (Prut and Belzung, 2003), questioning its suitability to capture anxiety behaviour. Moreover, most animal experimental designs tend to measure novelty-induced anxiety instead of long-term anxiety, which would be closer to the clinical setting. It has been suggested that the discrepancy can also be explained by a dissociation of social and generalized anxiety (Liu and Smith, 2009). Indeed, social anxiety is well documented in human literature; however, in preclinical studies it is confounded with other behavioural outcomes (*e.g.,* sociability) therefore, it was not possible to assess the fitness of the *Fmr1* KO model for this specific construct. However, while social phobia is the most common form of anxiety in FXS patients (Cordeiro et al., 2011), 50% of the patients show also generalized anxiety and 40% of the patients show agoraphobia, for which the open field test could be considered a very suitable test. Dissociation of generalized and social phobia can therefore only partly explain the discrepancies in anxiety phenotypes. Contrary to anxiety, the assessment of acoustic startle and PPI has a greater

level of similarity between species; however, the relevance of the auditory stimuli might differ between the species as they primarily rely on different senses. Compensatory upregulation of FMRP-associated proteins in the KO mice may underlie the opposite phenotypes (Frankland et al., 2004; Paylor et al., 2008), as double mutant mice lacking both *Fmr1* and *FXR2* (*FMR1* autosomal homolog 2) show decreased levels of PPI (Spencer et al., 2006). Furthermore, for all phenotypes which do not match the clinical profile it is important to keep in mind that these differences could be the result of a mismatch in disease induction in the KO models and patients. In contrast to the human condition, in neither of these two KO models the loss of protein is induced via an increase in CGG repeats. As the hypermethylation and thus silencing of protein expression in patients was shown to happen only at approximately the 12th day of gestation (Willemsen et al., 2002), differences in protein expression during early development could cause potential differences between the models and the clinical population. *Fmr1* knock-in (KI) models with increased CCG repeat expansions have been developed (Bontekoe et al., 2001; Entezam et al., 2007), but are currently only used to study the premutation (55-200 repeats) associated with Fragile X Tremor and Ataxia Syndrome (FXTAS). Although the mice also show repeat instability and permutation expansions that can develop into full mutation expansion numbers (>200 repeats; Entezam et al., 2007), for unknown reasons these expansion numbers are not resulting in protein silencing in mice (Entezam et al., 2007; Zhao et al., 2019). Therefore, the KO models are currently the best option to study FXS. However, in view of construct validity future studies should also consider to unravel why full mutation expansion numbers do not lead to protein silencing in mice in order to overcome the hurdles in developing functional KI models with increased CCG repeat expansions.

To be able to use anxiety, startle response and PPI in therapeutic interventions, it is important to further understand the phenotype discrepancies to allow for better interpretation and translation of rodent findings to clinical predictions.

In addition to assessing overall genotype effects, an important goal of this meta-analysis was to gain insight into factors that could explain the heterogeneity of the results in literature. Most of the overall genotypic effects scored a heterogeneity >70%, indicating high variability of the genotype effect between studies. This was also suggested by the substantial percentage of studies that reported a different direction of the effect than the overall effect.

Overall, few significant subgroup effects were found which only changed effect sizes but not the direction of effects. Additionally, the heterogeneity of the meta-analyses as assessed by the I^2-value, did not decrease after performing the subgroup analyses.

This includes the sex of the animal and the maternal genotype, which were expected to explain some of the variation based on the fact that FXS is an X-linked syndrome and earlier research showing differences between WT animals from WT or heterozygous dams (Zupan et al., 2016; Zupan and Toth, 2008). Together, these findings suggest that overlooked experimental factors introduced variability to our results. We speculate that the light phase in which the animals were tested and whether animals were single or group-housed could be relevant given their biological significance. These factors were included in the characteristics' extraction, but were reported too infrequently to be able to test their effects. Reporting these details information is important, as for example enriched environments have shown to reverse some of the phenotypes (Li et al., 2020; Restivo et al., 2005). In addition, we would like to highlight the infeasibility of making a cross-species assessment given the low number of studies performed with rats. Furthermore, it is possible that the variety of behavioural tests used explains part of the heterogeneity. Tests could differ in sensitivity to pick up certain phenotypes, or they may assess different aspects of the same phenotype. Although an exploratory analysis for this hypothesis did not show any indication of differences between the various tests used in the category of anxiety, our dataset allows for this assessment also in the other behavioural categories. These future analyses could not only give insight into whether different tests might pick up subtly different phenotypes, but also whether the between-study heterogeneity differs between the various behavioural tests available. Possibly, few effects were found as the assessed experimental factors do not affect the genotype effect independently, but interact among each other. Although the current analysis did not allow for assessing these interactions, current developments in complex modelling and machine learning would allow for extracting more information from the same data.

All in all, these results urgently call the preclinical research community to improve research practices and reporting to boost the quality of data to generate more meaningful and conclusive results; which also applies to systematic reviews and meta-analyses of preclinical studies (Hunniford et al., 2021). Since environmental factors are such a big driver of phenotypic variability, better reporting of experimental conditions is necessary to increase understanding of the true heterogeneity in results.

While systematic reviews and meta-analyses give comprehensive summaries of the existing literature, it is important to realize that they are not completely bias-free, as results of this meta-analysis are inherently dependent on the methodological decisions made to align the diverse datasets. For example, when due to phenotypic flexibility acoustic startle phenotypes may present themselves in different startle intensities across various experimental conditions, averaging within each study over all tested startle intensities

could lead to an underestimation of effect sizes. Unfortunately, as only a single outcome can be extracted per study, these kinds of decisions are unavoidable, nevertheless all decisions made for the current study were carefully discussed to minimize any kind of bias that could mislead the interpretation of results. However, some methodological decisions, in particular the boundaries of the age categories, are rather arbitrary since there is no consensus on these thresholds in literature. There are also studies stating that mice reach adulthood only after three months of age (Flurkey et al., 2007), and age subgroup analysis using this threshold actually showed a significant age effect on the anxiety phenotype (data not shown). Because of the non-consensus about these thresholds in the field, reporting the actual ages should always be preferred over only reporting the developmental stage of the experimental animals. An additional limitation of meta-analyses is that there are currently no automated methods to perform the data and characteristics extraction, therefore they are prone to human error. However, when running a random sub-sample check we only found 3.5% of errors in the characteristics extraction; given the large size of the meta-analyses, these errors minimally altered the effect sizes and did not change any of the outcomes of subgroup analysis (i.e., conclusions stayed the same).

Additionally, the quality of this meta-analysis is dependent on the quality of the data included. There are numerous accounts, including the quality assessment in this meta-analysis, highlighting the often poor reporting and flawed experimental design of many preclinical studies (Kilkenny et al., 2009; Landis et al., 2012). This includes the prevalent lack of use, or reporting, of blinding and randomization, and poor research practices such as inadequate statistical tests and p-HARking (formulating the Hypothesis After gathering Results) (Bishop, 2019). The risk of bias assessment of the current systematic review and meta-analysis showed suboptimal reporting. Only 48.9% of the screened studies reported being 'blinded' for the experimental groups when conducting experiments while only 59.5% reported it for assessing the outcome. In 86.7% of studies, the housing arrangement of the animal subjects was unclear *(e.g.,* group housed with mixed genotypes). Strikingly, 52.3% did not report the baseline characteristics of their experimental samples: a) whether the control and experimental group were littermates, b) age and c) gender of the animal subjects. The poor reporting of experimental details also hindered subgroup analysis. Lastly, we were bound by the amount of information available, there might be more data available which was never published. There are indications that preclinical studies overestimate the treatment effectiveness by 30% partly due to the absence of published neutral results (*i.e.*, non-significant) and lack of methodological rigor (Sena et al., 2014). A publication bias analysis indicated missing data in multiple categories, and trim and fill analysis showed that the effects on acoustic

startle and repetitive behaviour were no longer significant after imputing missing studies, highlighting the consequences publication bias can have.

Taken together, this systematic review and meta-analysis show that the robustness as well as translatability of the *Fmr1* KO model to the clinical profile varies over the different behavioural phenotypes. Overall, many significant phenotypes were found with the same effect direction as seen in patients, thus showing good translational validity. However, altogether there was a large heterogeneity between studies and many effect sizes were relatively small. For most phenotypes there was low replicability which, despite translational validity, asks for careful interpretation of individual study findings. Additionally, when designing a study where the use of the *Fmr1* KO model is considered, one should be aware of the not fully understood mismatch in rodent and clinical phenotypes (*e.g.,* anxiety, startle and PPI, aggression, attention and cognitive flexibility). The cognitive and audiogenic seizure phenotypes showed the highest replicability, in addition to translational validity, therefore the intellectual disability and epilepsy elements of FXS are possibly the most meaningful to study with the *Fmr1* KO. The model as a whole should be more cautiously used for the ASD-like elements of the disorder, which showed translational validity, but the replicability of these phenotypes was low. More importantly, the phenotypic and quality results provided by this meta-analysis urge for a broad reappraisal of the current research and reporting practices in all preclinical models of brain disorders to deliver more meaningful preclinical data.

ACKNOWLEDGEMENTS

This project has received funding from a ZonMW TOP subsidy (grant number 91216021, 2017) and a MKMD subsidy (grant number 114024154, 2020). In addition, this project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777364. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

AUTHOR CONTRIBUTIONS

R. Kat: Conceptualization, funding acquisition, investigation, formal analysis, visualization, writing - original draft; **M. Arroyo-Araujo:** Investigation, formal analysis, visualization, writing - original draft; R. de Vries: Conceptualization, supervision, writing - reviewing and editing; **M.A. Koopmans:** Investigation; **S.F. de Boer:** Conceptualization, supervision, writing - reviewing and editing; **M.J.H. Kas:** Conceptualization, supervision, funding acquisition, writing - reviewing and editing)

REFERENCES

Adriani, W., Granstrem, O., Macri, S., Izykenova, G., Dambinova, S., Laviola, G., 2004. Behavioral and neurochemical vulnerability during adolescence in mice: Studies with nicotine. Neuropsychopharmacology 29, 869–878. https://doi.org/10.1038/sj.npp.1300366

Adusei, D.C., Pacey, L.K.K., Chen, D., Hampson, D.R., 2010. Early developmental alterations in GABAergic protein expression in fragile X knockout mice. Neuropharmacology 59, 167–171. https://doi.org/10.1016/j.neuropharm.2010.05.002

Ahnaou, A., Moechars, D., Raeymaekers, L., Biermans, R., Manyakov, N. V., Bottelbergs, A., Wintmolders, C., Van Kolen, K., Van De Casteele, T., Kemp, J.A., Drinkenburg, W.H., 2017. Emergence of early alterations in network oscillations and functional connectivity in a tau seeding mouse model of Alzheimer's disease pathology. Sci. Rep. 7, 1–14. https://doi.org/10.1038/s41598-017-13839-6

Alam, R.U., Zhao, H., Goodwin, A., Kavehei, O., McEwan, A., 2020. Differences in power spectral densities and phase quantities due to processing of eeg signals. Sensors (Switzerland) 20, 1–20. https://doi.org/10.3390/s20216285

Allen Institute for Brain Science, 2014. Adult human reference atlas modified Brodmann [WWW Document]. URL https://atlas.brain-map.org/

Allen Institute for Brain Science, 2011. Adult mouse reference atlas saggital [WWW Document]. URL https://atlas.brain-map.org/

Aman, M.G., 2012. Aberrant behavior checklist: Current identity and future developments. Clin. Exp. Pharmacol. 2, 357–373. https://doi.org/10.4172/2161-1459.1000e114

Anderson, A., Locke, J., Kretzmann, M., Kasari, C., 2016. Social network analysis of children with autism spectrum disorder: Predictors of fragmentation and connectivity in elementary school classrooms. Autism 20, 700–709. https://doi.org/10.1177/1362361315603568

Angelakos, C.C., Tudor, J.C., Ferri, S.L., Jongens, T.A., Abel, T., 2019. Home-cage hypoactivity in mouse genetic models of autism spectrum disorder. Neurobiol. Learn. Mem. 165, 0–1. https://doi.org/10.1016/j.nlm.2019.02.010

Antoine, M.W., Langberg, T., Schnepel, P., Feldman, D.E., 2019. Increased Excitation-Inhibition Ratio Stabilizes Synapse and Circuit Excitability in Four Autism Mouse Models. Neuron 101, 648-661.e4. https://doi.org/10.1016/j.neuron.2018.12.026

Assaf, Y., Bouznach, A., Zomet, O., Marom, A., Yovel, Y., 2020. Conservation of brain connectivity and wiring across the mammalian class. Nat. Neurosci. 23, 805–808. https://doi.org/10.1038/s41593-020-0641-7

Avey, M.T., Moher, D., Sullivan, K.J., Fergusson, D., Griffin, G., Grimshaw, J.M., Hutton, B., Lalu, M.M., Macleod, M., Marshall, J., Mei, S.H.J., Rudnicki, M., Stewart, D.J., Turgeon, A.F., McIntyre, L., 2016. The devil is in the details: Incomplete reporting in preclinical animal research. PLoS One 11, 1–13. https://doi.org/10.1371/journal.pone.0166733

Bagni, C., Oostra, B.A., 2013. Fragile X syndrome: From protein function to therapy. Am. J. Med. Genet. Part A 161, 2809–2821. https://doi.org/10.1002/ajmg.a.36241

Bailey, D.B., Mesibov, G.B., Hatton, D.D., Clark, R.D., Roberts, J.E., Mayhew, L., 1998. Autistic behavior in young boys with fragile X syndrome. J. Autism Dev. Disord. 28, 499–508. https://doi.org/10.1023/A:1026048027397

Baker, K.B., Wray, S.P., Ritter, R., Mason, S., Lanthorn, T.H., Savelieva, K. V, 2010. Male and female Fmr1 knockout mice on C57 albino background exhibit spatial learning and memory impairments. Genes. Brain. Behav. 9, 562–574. https://doi.org/10.1111/j.1601-183X.2010.00585.x

Baker, M., 2013. Neuroscience: Through the eyes of a mouse. Nature 502, 156–158. https://doi.org/10.1038/502156a

Baranek, G.T., Chin, Y.H., Greiss Hess, L.M., Yankee, J.G., Hatton, D.D., Hooper, S.R., 2002. Sensory processing correlates of occupational performance in children with fragile X syndrome: Preliminary findings. Am. J. Occup. Ther. 56, 538–546. https://doi.org/10.5014/ajot.56.5.538

Baranek, G.T., Roberts, J.E., David, F.J., Sideris, J., Penny, L., Hatton, D.D., Bailey, D.B., Baranek, G.T., Roberts, J.E., David, F.J., Sideris, J., Penny, L., Hatton, D.D., Bailey, D.B., Trajectories, D., Roberts, J.E., David, F.J., Mirrett, P.L., Hatton, D.D., Bailey, D.B., 2009. Developmental Trajectories and Correlates of Sensory Processing in Young Boys with Fragile X Syndrome Developmental Trajectories and Correlates of Sensory Processing in Young Boys with Fragile X Syndrome. Phys. Occup. Ther. Pediatr. 28, 79–98. https://doi.org/10.1300/J006v28n01_06

Barral, J., D'Reyes, A., 2016. Synaptic scaling rule preserves excitatory-inhibitory balance and salient neuronal network dynamics. Nat. Neurosci. 19, 1690–1696. https://doi.org/10.1038/nn.4415

Basu, S.N., Kollu, R., Banerjee-Basu, S., 2009. AutDB: A gene reference resource for autism research. Nucleic Acids Res. 37, 832–836. https://doi.org/10.1093/nar/gkn835

Bates, D., Kliegl, R., Vasishth, S., Baayen, H., 2015a. Parsimonious Mixed Models. ArXiv e-print 1506.04967v2.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015b. Package lme4. J. Stat. Softw. 67, 1–91. https://doi.org/http://lme4.r-forge.r-project.org

Baum, S.H., Stevenson, R.A., Wallace, M.T., 2015. Behavioral, perceptual, and neural alterations in sensory and multisensory function in autism spectrum disorder. Prog. Neurobiol. 134, 140–160. https://doi.org/10.1016/j.pneurobio.2015.09.007

Baumgardner, T.L., Reiss, A.L., Freund, L.S., Abrams, M.T., 1995. Specification of the neurobehavioral phenotype in males with fragile X syndrome. Pediatrics 95, 744–752.

Belzung, C., Lemoine, M., 2011. Criteria of validity for animal models of psychiatric disorders: focus on anxiety disorders and depression. Biol. Mood Anxiety Disord. 1, 1–14. https://doi.org/10.1186/2045-5380-1-9

Ben-Ari, Y., Khalilov, I., Kahle, K.T., Cherubini, E., 2012. The GABA excitatory/inhibitory shift in brain maturation and neurological disorders. Neuroscientist 18, 467–486. https://doi.org/10.1177/1073858412438697

Berry-Kravis, E., Hessl, D., Coffey, S., Hervey, C., Schneider, A., Yuhas, J., Hutchison, J., Snape, M., Tranfaglia, M., Nguyen, D. V., Hagerman, R., 2009. A pilot open label, single dose trial of fenobam in adults with fragile X syndrome. J. Med. Genet. 46, 266–271. https://doi.org/10.1136/jmg.2008.063701

Berry-Kravis, E.M., Lindemann, L., Jønch, A.E., Apostol, G., Bear, M.F., Carpenter, R.L., Crawley, J.N., Curie, A., Des Portes, V., Hossain, F., Gasparini, F., Gomez-Mancilla, B., Hessl, D., Loth, E., Scharf, S.H., Wang, P.P., Von Raison, F., Hagerman, R., Spooren, W., Jacquemont, S., 2018. Drug development for neurodevelopmental disorders: Lessons learned from fragile X syndrome. Nat. Rev. Drug Discov. 17, 280–298. https://doi.org/10.1038/nrd.2017.221

Berzhanskaya, J., Phillips, M.A., Shen, J., Colonnese, M.T., 2016. Sensory hypo-excitability in a rat model of fetal development in Fragile X Syndrome. Sci. Rep. 6, 1–11. https://doi.org/10.1038/srep30769

Bespalov, A., Bernard, R., Gilis, A., Gerlach, B., Guillén, J., Castagné, V., Lefevre, I.A., Ducrey, F., Monk, L., Bongiovanni, S., Altevogt, B., Arroyo-Araujo, M., Bikovski, L., de Bruin, N., Castaños-Vélez, E., Dityatev, A., Emmerich, C.H., Fares, R., Ferland-Beckham, C., Froger-Colléaux, C., Gailus-Durner, V., Hölter, S.M., Hofmann, M.C.J., Kabitzke, P., Kas, M.J.H., Kurreck, C., Moser, P., Pietraszek, M., Popik, P., Potschka, H., de Oca, E.P.M., Restivo, L., Riedel, G., Ritskes-Hoitinga, M., Samardzic, J., Schunn, M., Stöger, C., Voikar, V., Vollert, J., Wever, K.E., Wuyts, K., Macleod, M.R., Dirnagl, U., Steckler, T., 2021. Introduction to the eqipd quality system. Elife 10, 1–26. https://doi.org/10.7554/eLife.63294

Bespalov, A., Steckler, T., 2018. Lacking quality in research: Is behavioral neuroscience affected more than other areas of biomedical science? J. Neurosci. Methods 300, 4–9. https://doi.org/10.1016/j.jneumeth.2017.10.018

Bey, A.L., Wang, X., Yan, H., Kim, N., Passman, R.L., Yang, Y., Cao, X., Towers, A.J., Hulbert, S.W., Duffney, L.J., Gaidis, E., Rodriguiz, R.M., Wetsel, W.C., Yin, H.H., Jiang, Y.H., 2018. Brain region-specific disruption of Shank3 in mice reveals a dissociation for cortical and striatal circuits in autism-related behaviors. Transl. Psychiatry 8. https://doi.org/10.1038/s41398-018-0142-6

Bhattacharya, A., Mamcarz, M., Mullins, C., Choudhury, A., Boyle, R.G., Smith, D.G., Walker, D.W., Klann, E., 2016. Targeting Translation Control with p70 S6 Kinase 1 Inhibitors to Reverse Phenotypes in Fragile X Syndrome Mice. Neuropsychopharmacology 41, 1991–2000. https://doi.org/10.1038/npp.2015.369

Bishop, D., 2019. Rein in the four horsemen of irreproducibility. Nature 568, 435. https://doi.org/10.1038/d41586-019-01307-2

Bodden, C., von Kortzfleisch, V.T., Karwinkel, F., Kaiser, S., Sachser, N., Richter, S.H., 2019. Heterogenising study samples across testing time improves reproducibility of behavioural data. Sci. Rep. 9, 1–9. https://doi.org/10.1038/s41598-019-44705-2

Bontekoe, C.J.M., Bakker, C.E., Nieuwenhuizen, I.M., Van Der Linde, H., Lans, H., De Lange, D., Hirst, M.C., Oostra, B.A., 2001. Instability of a (CGG)98 repeat in the Fmr1 promoter. Hum. Mol. Genet. 10, 1693–1699. https://doi.org/10.1093/hmg/10.16.1693

Boucher, M.N., Aktar, M., Braas, K.M., May, V., Hammack, S.E., 2021. Activation of Lateral Parabrachial Nucleus (LPBn) PACAP - Expressing Projection Neurons to the Bed Nucleus of the Stria Terminalis (BNST) Enhances Anxiety - like Behavior. J. Mol. Neurosci. https://doi.org/10.1007/s12031-021-01946-z

Braat, S., D'Hulst, C., Heulens, I., de Rubeis, S., Mientjes, E., Nelson, D.L., Willemsen, R., Bagni, C., van Dam, D., de Deyn, P.P., Kooy, R.F., 2015. The GABAA receptor is an FMRP target with therapeutic potential in fragile X syndrome. Cell Cycle 14, 2985–2995. https://doi.org/10.4161/15384101.2014.989114

Bruining, H., Hardstone, R., Juarez-Martinez, E.L., Sprengers, J., Avramiea, A.E., Simpraga, S., Houtman, S.J., Poil, S.S., Dallares, E., Palva, S., Oranje, B., Matias Palva, J., Mansvelder, H.D., Linkenkaer-Hansen, K., 2020. Measurement of excitation-inhibition ratio in autism spectrum disorder using critical brain dynamics. Sci. Rep. 10, 9195. https://doi.org/10.1038/s41598-020-65500-4

Bruining, H., Passtoors, L., Goriounova, N., Jansen, F., Hakvoort, B., de Jonge, M., Poil, S.-S., 2015. Paradoxical Benzodiazepine Response: A Rationale for Bumetanide in Neurodevelopmental Disorders? Pediatrics 136, e539–e543. https://doi.org/10.1542/peds.2014-4133

Building a better mouse test, 2011. . Nat. Methods 8, 697. https://doi.org/10.1038/nmeth.1698

Çaku, A., Pellerin, D., Bouvier, P., Riou, E., Corbin, F., 2014. Effect of lovastatin on behavior in children and adults with fragile X syndrome: An open-label study. Am. J. Med. Genet. Part A 164, 2834–2842. https://doi.org/10.1002/ajmg.a.36750

Campbell, D.T., Stanley, J.C., 1963. Experimental and quasi-experimental designs for research. Houghton Mifflin Company, Boston, USA. https://doi.org/10.1037/022808

Carrillo-Reid, L., Han, S., Yang, W., Akrouh, A., Yuste, R., 2019. Controlling Visually Guided Behavior by Holographic Recalling of Cortical Ensembles. Cell 178, 447-457.e5. https://doi.org/10.1016/j.cell.2019.05.045

Castrén, M., Pääkkönen, A., Tarkka, I.M., Ryynänen, M., Partanen, J., 2003. Augmentation of auditory N1 in children with fragile X syndrome. Brain Topogr. 15, 165–171. https://doi.org/10.1023/A:1022606200636

Chen, C.C., Kiebel, S.J., Kilner, J.M., Ward, N.S., Stephan, K.E., Wang, W.J., Friston, K.J., 2012. A dynamic causal model for evoked and induced responses. Neuroimage 59, 340–348. https://doi.org/10.1016/j.neuroimage.2011.07.066

Chen, Q., Deister, C.A., Gao, X., Guo, B., Lynn-Jones, T., Chen, N., Wells, M.F., Liu, R., Goard, M.J., Dimidschstein, J., Feng, S., Shi, Y., Liao, W., Lu, Z., Fishell, G., Moore, C.I., Feng, G., 2020. Dysfunction of cortical GABAergic neurons leads to sensory hyper-reactivity in a Shank3 mouse model of ASD. Nat. Neurosci. 23, 520–532. https://doi.org/10.1038/s41593-020-0598-6

Chen, W., Cai, Z.L., Chao, E.S., Chen, H., Longley, C.M., Hao, S., Chao, H.T., Kim, J.H., Messier, J.E., Zoghbi, H.Y., Tang, J., Swann, J.W., Xue, M., 2020. Stxbp1/Munc18-1 haploinsufficiency impairs inhibition and mediates key neurological features of STXBP1 encephalopathy. Elife 9, 1–33. https://doi.org/10.7554/eLife.48705

Chen, X., Tong, C., Han, Z., Zhang, K., Bo, B., Feng, Y., Liang, Z., 2020. Sensory evoked fMRI paradigms in awake mice. Neuroimage 204, 116242. https://doi.org/10.1016/j.neuroimage.2019.116242

Cheval, H., Guy, J., Merusi, C., De Sousa, D., Selfridge, J., Bird, A., 2012. Postnatal inactivation reveals enhanced requirement for MeCP2 at distinct age windows. Hum. Mol. Genet. 21, 3806–3814. https://doi.org/10.1093/hmg/dds208

Cheyne, J.E., Zabouri, N., Baddeley, D., Lohmann, C., 2019. Spontaneous Activity Patterns Are Altered in the Developing Visual Cortex of the Fmr1 Knockout Mouse. Front. Neural Circuits 13, 1–8. https://doi.org/10.3389/fncir.2019.00057

Cho, K.K.A., Hoch, R., Lee, A.T., Patel, T., Rubenstein, J.L.R., Sohal, V.S., 2015. Gamma rhythms link prefrontal interneuron dysfunction with cognitive inflexibility in dlx5/6+/- mice. Neuron 85, 1332–1343. https://doi.org/10.1016/j.neuron.2015.02.019

Ciaccio, C., Fontana, L., Milani, D., Tabano, S., Miozzo, M., Esposito, S., 2017. Fragile X syndrome: a review of clinical and molecular diagnoses. Ital. J. Pediatr. 43, 1–12. https://doi.org/10.1186/s13052-017-0355-y

Cirrito, J.R., Wallace, C.E., Yan, P., Davis, T.A., Gardiner, W.D., Doherty, B.M., King, D., Yuede, C.M., Lee, J.M., Sheline, Y.I., 2020. Effect of escitalopram on Aβ levels and plaque load in an Alzheimer mouse model. Neurology 95, e2666–e2674. https://doi.org/10.1212/WNL.0000000000010733

Clapp, W.C., Eckert, M.J., Teyler, T.J., Abraham, W.C., 2006. Rapid visual stimulation induces N-methyl-D-aspartate receptor-dependent sensory long-term potentiation in the rat cortex. Neuroreport 17, 511–515. https://doi.org/10.1097/01.wnr.0000209004.63352.10

Cogram, P., Alkon, D.L., Crockford, D., Deacon, R.M.J., Hurley, M.J., Altimiras, F., Sun, M.K., Tranfaglia, M., 2020. Chronic bryostatin-1 rescues autistic and cognitive phenotypes in the fragile X mice. Sci. Rep. 10, 1–10. https://doi.org/10.1038/s41598-020-74848-6

Cogram, P., Deacon, R.M.J., Warner-Schmidt, J.L., von Schimmelmann, M.J., Abrahams, B.S., During, M.J., 2019. Gaboxadol normalizes behavioral abnormalities in a mouse model of fragile x syndrome. Front. Behav. Neurosci. 13, 1–9. https://doi.org/10.3389/fnbeh.2019.00141

Cohen, M.X., 2014. Analyzing neural time series data: theory and practice, 1st ed. The MIT press, Cambridge, Massachusetts.

Cordeiro, L., Ballinger, E., Hagerman, R., Hessl, D., 2011. Clinical assessment of DSM-IV anxiety disorders in fragile X syndrome: prevalence and characterization. J. Neurodev. Disord. 3, 57–67. https://doi.org/10.1007/s11689-010-9067-y

Crescitelli, F., Gardner, E., 1961. Correspondences in the behavior of the electroretinogram and of the potentials evoked at the visual cortex. J. Gen. Physiol. 44, 911–928. https://doi.org/10.1085/jgp.44.5.911

Crusio, W.E., 2015. Key issues in contemporary behavioral genetics. Curr. Opin. Behav. Sci. 2, 89–95. https://doi.org/10.1016/j.cobeha.2014.10.002

Czigler, I., 2007. Visual mismatch negativity: Violation of nonattended environmental regularities. J. Psychophysiol. 21, 224–230. https://doi.org/10.1027/0269-8803.21.34.224

Czigler, I., Weisz, J., Winkler, I., 2006. ERPs and deviance detection: Visual mismatch negativity to repeated visual stimuli. Neurosci. Lett. 401, 178–182. https://doi.org/10.1016/j.neulet.2006.03.018

D'Hulst, C., De Geest, N., Reeve, S.P., Van Dam, D., De Deyn, P.P., Hassan, B.A., Kooy, R.F., 2006. Decreased expression of the GABAA receptor in fragile X syndrome. Brain Res. 1121, 238–245. https://doi.org/10.1016/j.brainres.2006.08.115

Dahlhaus, R., El-Husseini, A., 2010. Altered neuroligin expression is involved in social deficits in a mouse model of the fragile X syndrome. Behav. Brain Res. 208, 96–105. https://doi.org/10.1016/j.bbr.2009.11.019

Deacon, R.M.J., Glass, L., Snape, M., Hurley, M.J., Altimiras, F.J., Biekofsky, R.R., Cogram, P., 2015. NNZ-2566, a novel analog of (1-3) IGF-1, as a potential therapeutic agent for fragile X syndrome. Neuromolecular Med. 17, 71–82. https://doi.org/10.1007/s12017-015-8341-2

den Broeder, M.J., van der Linde, H., Brouwer, J.R., Oostra, B.A., Willemsen, R., Ketting, R.F., 2009. Generation and characterization of Fmr1 knockout zebrafish. PLoS One 4, 2–7. https://doi.org/10.1371/journal.pone.0007910

Deng, P.Y., Sojka, D., Klyachko, V.A., 2011. Abnormal presynaptic short-term plasticity and information processing in a mouse model of fragile X syndrome. J. Neurosci. 31, 10971–10982. https://doi.org/10.1523/JNEUROSCI.2021-11.2011

Dickinson, A., Jones, M., Milne, E., 2016. Measuring neural excitation and inhibition in autism: Different approaches, different findings and different interpretations. Brain Res. 1648, 277–289. https://doi.org/10.1016/j.brainres.2016.07.011

Ding, Q., Sethna, F., Wang, H., 2014. Behavioral analysis of male and female Fmr1 knockout mice on C57BL/6 background. Behav. Brain Res. 271, 72–78. https://doi.org/10.1016/j.bbr.2014.05.046

Ding, Q., Sethna, F., Wu, X.T., Miao, Z., Chen, P., Zhang, Y., Xiao, H., Feng, W., Feng, Y., Li, X., Wang, H., 2020. Transcriptome signature analysis repurposes trifluoperazine for the treatment of fragile X syndrome in mouse model. Commun. Biol. 3, 1–13. https://doi.org/10.1038/s42003-020-0833-4

Dionne, O., Corbin, F., 2021. An "omic" overview of fragile X syndrome. Biology (Basel). 10. https://doi.org/10.3390/biology10050433

Duval, S., Tweedie, R., 2000. Trim and Fill: A Simple Funnel-Plot-Based Method. Biometrics 56, 455–463.

Edwards, D.J., Trujillo, L.T., 2021. An analysis of the external validity of eeg spectral power in an uncontrolled outdoor environment during default and complex neurocognitive states. Brain Sci. 11, 1–26. https://doi.org/10.3390/brainsci11030330

Egger, M., Smith, G.D., Schneider, M., Minder, C., 1997. Bias in meta-analysis detected by a simple, graphical test. Br. Med. J. 315, 629–634. https://doi.org/10.1136/bmj.315.7109.629

Entezam, A., Biacsi, R., Orrison, B., Saha, T., Hoffman, G.E., Grabczyk, E., Nussbaum, R.L., Usdin, K., 2007. Regional FMRP deficits and large repeat expansions into the full mutation range in a new Fragile X premutation mouse model. Gene 395, 125–134. https://doi.org/10.1016/j.gene.2007.02.026

Ethridge, L.E., De Stefano, L.A., Schmitt, L.M., Woodruff, N.E., Brown, K.L., Tran, M., Wang, J., Pedapati, E. V., Erickson, C.A., Sweeney, J.A., 2019. Auditory EEG Biomarkers in Fragile X Syndrome: Clinical Relevance. Front. Integr. Neurosci. 13, 1–16. https://doi.org/10.3389/fnint.2019.00060

Ethridge, L.E., White, S.P., Mosconi, M.W., Wang, J., Byerly, M.J., Sweeney, J.A., 2016. Reduced habituation of auditory evoked potentials indicate cortical hyper-excitability in Fragile X Syndrome. Transl. Psychiatry 6, e787. https://doi.org/10.1038/tp.2016.48

Ethridge, L.E., White, S.P., Mosconi, M.W., Wang, J., Pedapati, E. V, Erickson, C.A., Byerly, M.J., Sweeney, J.A., 2017. Neural synchronization deficits linked to cortical hyper-excitability and auditory hypersensitivity in fragile X syndrome. Mol. Autism 8, 22. https://doi.org/10.1186/s13229-017-0140-1

Featherstone, R.E., Shin, R., Kogan, J.H., Liang, Y., Matsumoto, M., Siegel, S.J., 2015. Mice with subtle reduction of NMDA NR1 receptor subunit expression have a selective decrease in mismatch negativity: Implications for schizophrenia prodromal population. Neurobiol. Dis. 73, 289–295. https://doi.org/10.1016/j.nbd.2014.10.010

Felgerolle, C., Hébert, B., Ardourel, M., Meyer-Dilhet, G., Menuet, A., Pinto-Morais, K., Bizot, J.C., Pichon, J., Briault, S., Perche, O., 2019. Visual Behavior Impairments as an Aberrant Sensory Processing in the Mouse Model of Fragile X Syndrome. Front. Behav. Neurosci. 13, 1–19. https://doi.org/10.3389/fnbeh.2019.00228

File, D., File, B., Bodnár, F., Sulykos, I., Kecskés-Kovács, K., Czigler, I., 2017. Visual mismatch negativity (vMMN) for low- and high-level deviances: A control study. Attention, Perception, Psychophys. 79, 2153–2170. https://doi.org/10.3758/s13414-017-1373-y

Flurkey, K., Currer, J.M., Harrison, D.E., 2007. Mouse Models in Aging Research, in: The Mouse in Biomedical Research. Academic Press, Cambridge, Massachusetts, pp. 637–672.

Fournier, J., Saleem, A.B., Diamanti, E.M., Wells, M.J., Harris, K.D., Carandini, M., 2020. Mouse Visual Cortex Is Modulated by Distance Traveled and by Theta Oscillations. Curr. Biol. 30, 3811-3817.e6. https://doi.org/10.1016/j.cub.2020.07.006

Frankland, P.W., Wang, Y., Rosner, B., Shimizu, T., Balleine, B.W., Dykens, E.M., Ornitz, E.M., Silva, A.J., 2004. Sensorimotor gating abnormalities in young males with fragile X syndrome and Fmr1-knockout mice. Mol. Psychiatry 9, 417–425. https://doi.org/10.1038/sj.mp.4001432

Gantois, I., Khoutorsky, A., Popic, J., Aguilar-Valles, A., Freemantle, E., Cao, R., Sharma, V., Pooters, T., Nagpal, A., Skalecka, A., Truong, V.T., Wiebe, S., Groves, I.A., Jafarnejad, S.M., Chapat, C., McCullagh, E.A., Gamache, K., Nader, K., Lacaille, J.-C., Gkogkas, C.G., Sonenberg, N., 2017. Metformin ameliorates core deficits in a mouse model of fragile X syndrome. Nat. Med. 23, 674–677. https://doi.org/10.1038/nm.4335

Gantois, I., Pop, A.S., de Esch, C.E.F., Buijsen, R.A.M., Pooters, T., Gomez-Mancilla, B., Gasparini, F., Oostra, B.A., D'Hooge, R., Willemsen, R., 2013. Chronic administration of AFQ056/Mavoglurant restores social behaviour in Fmr1 knockout mice. Behav. Brain Res. 239, 72–79. https://doi.org/10.1016/j.bbr.2012.10.059

Garland, T., Kelly, S.A., 2006. Phenotypic plasticity and experimental evolution. J. Exp. Biol. 209, 2344–2361. https://doi.org/10.1242/jeb.02244

Garrido, M.I., Kilner, J.M., Stephan, K.E., Friston, K.J., 2009. The mismatch negativity: A review of underlying mechanisms. Clin. Neurophysiol. 120, 453–463. https://doi.org/10.1016/j.clinph.2008.11.029

Gaudissard, J., Ginger, M., Premoli, M., Memo, M., Frick, A., Pietropaolo, S., 2017. Behavioral abnormalities in the Fmr1-KO2 mouse model of fragile X syndrome: The relevance of early life phases. Autism Res. 10, 1584–1596. https://doi.org/10.1002/aur.1814

Gelman, a, Hill, J., 2007. Data analysis using regression and multilevel/hierarchical models. Policy Anal. 1–651. https://doi.org/10.2277/0521867061

Ghasemi, Asghar; Sajad, J.K.K., 2021. The laboratory rat: age and body weight matter. EXCLI J. 20, 1431–1445.

Giordano, G.M., Brando, F., Perrottelli, A., Di Lorenzo, G., Siracusano, A., Giuliani, L., Pezzella, P., Altamura, M., Bellomo, A., Cascino, G., Del Casale, A., Monteleone, P., Pompili, M., Galderisi, S., Maj, M., 2021. Tracing Links Between Early Auditory Information Processing and Negative Symptoms in Schizophrenia: An ERP Study. Front. Psychiatry 12, 1–12. https://doi.org/10.3389/fpsyt.2021.790745

Gliske, S. V., Irwin, Z.T., Chestek, C., Stacey, W.C., 2016. Effect of sampling rate and filter settings on High Frequency Oscillation detections. Clin. Neurophysiol. 127, 3042–3050. https://doi.org/10.1016/j.clinph.2016.06.029

Goel, A., Cantu, D.A., Guilfoyle, J., Chaudhari, G.R., Newadkar, A., Todisco, B., de Alba, D., Kourdougli, N., Schmitt, L.M., Pedapati, E., Erickson, C.A., Portera-Cailliau, C., 2018. Impaired perceptual learning in a mouse model of Fragile X syndrome is mediated by parvalbumin neuron dysfunction and is reversible. Nat. Neurosci. 21, 1404–1411. https://doi.org/10.1038/s41593-018-0231-0

Gonçalves, J.T., Anstey, J.E., Golshani, P., Portera-Cailliau, C., 2013. Circuit level defects in the developing neocortex of Fragile X mice. Nat. Neurosci. 16, 903–909. https://doi.org/10.1038/nn.3415

Grimm, S., Escera, C., Nelken, I., 2016. Early indices of deviance detection in humans and animal models. Biol. Psychol. 116, 23–27. https://doi.org/10.1016/j.biopsycho.2015.11.017

Gromer, D., Kiser, D.P., Pauli, P., 2021. Thigmotaxis in a virtual human open field test. Sci. Rep. 11, 1–13. https://doi.org/10.1038/s41598-021-85678-5

Gurney, M.E., Cogram, P., Deacon, R.M., Rex, C., Tranfaglia, M., 2017. Multiple Behavior Phenotypes of the Fragile-X Syndrome Mouse Model Respond to Chronic Inhibition of Phosphodiesterase-4D (PDE4D). Sci. Rep. 7, 14653. https://doi.org/10.1038/s41598-017-15028-x

Hagerman, R.J., Jackson, A.W., Levitas, A., Rimland, B., Braden, M., 1986. An analysis of autism in fifty males with the fragile X syndrome. Am. J. Med. Genet. 23, 359–374. https://doi.org/10.1002/ajmg.1320230128

Hamm, J.P., Shymkiv, Y., Mukai, J., Gogos, J.A., Yuste, R., 2020. Aberrant Cortical Ensembles and Schizophrenia-like Sensory Phenotypes in Setd1a+/− Mice. Biol. Psychiatry 88, 215–223. https://doi.org/10.1016/j.biopsych.2020.01.004

Hamm, J.P., Yuste, R., 2016. Somatostatin Interneurons Control a Key Component of Mismatch Negativity in Mouse Visual Cortex. Cell Rep. 16, 597–604. https://doi.org/10.1016/j.celrep.2016.06.037

Han, K., Chen, H., Gennarino, V.A., Richman, R., Lu, H.C., Zoghbi, H.Y., 2015. Fragile X-like behaviors and abnormal cortical dendritic spines in Cytoplasmic FMR1-interacting protein 2-mutant mice. Hum. Mol. Genet. 24, 1813–1823. https://doi.org/10.1093/hmg/ddu595

Hånell, A., Marklund, N., 2014. Structured evaluation of rodent behavioral tests used in drug discovery research. Front. Behav. Neurosci. 8, 1–13. https://doi.org/10.3389/fnbeh.2014.00252

Hansen, I.H., Agerskov, C., Arvastson, L., Bastlund, J.F., Sørensen, H.B.D., Herrik, K.F., 2019. Pharmaco-electroencephalographic responses in the rat differ between active and inactive locomotor states. Eur. J. Neurosci. 50, 1948–1971. https://doi.org/10.1111/ejn.14373

Hardstone, R., Poil, S.S., Schiavone, G., Jansen, R., Nikulin, V. V., Mansvelder, H.D., Linkenkaer-Hansen, K., 2012. Detrended fluctuation analysis: A scale-free view on neuronal oscillations. Front. Physiol. 3 NOV, 1–13. https://doi.org/10.3389/fphys.2012.00450

Harms, L., Michie, P.T., Näätänen, R., 2016. Criteria for determining whether mismatch responses exist in animal models: Focus on rodents. Biol. Psychol. 116, 28–35. https://doi.org/10.1016/j.biopsycho.2015.07.006

Harris, S.W., Hessl, D., Goodlin-Jones, B., Ferranti, J., Bacalman, S., Barbato, I., Tassone, F., Hagerman, P.J., Herman, K., Hagerman, R.J., 2008. Autism profiles of males with fragile X syndrome. Am. J. Ment. Retard. 113, 427–438. https://doi.org/10.1352/2008.113:427-438

He, C.X., Portera-Cailliau, C., 2013. The trouble with spines in fragile X syndrome: Density, maturity and plasticity. Neuroscience 251, 120–128. https://doi.org/10.1016/j.neuroscience.2012.03.049

He, Q., Nomura, T., Xu, J., Contractor, A., 2014. The Developmental Switch in GABA Polarity Is Delayed in Fragile X Mice. J. Neurosci. 34, 446–450. https://doi.org/10.1523/JNEUROSCI.4447-13.2014

Heffner, H., Masterton, B., 1980. Hearing in Glires: Domestic rabbit, cotton rat, feral house mouse, and kangaroo rat. J. Acoust. Soc. Am. 68, 1584–1599. https://doi.org/10.1121/1.385213

Heintz, T., Hinojosa, A., Lagnado, L., 2020. Opposing forms of adaptation in mouse visual cortex are controlled by distinct inhibitory microcircuits and gated by locomotion. BioRxiv. https://doi.org/10.1101/2020.01.16.909788

Hersh, J.H., Saul, R.A., Saal, H.M., Braddock, S.R., Enns, G.M., Gruen, J.R., Perrin, J.M., Tarini, B.A., 2011. Clinical report-health supervision for children with fragile X syndrome. Pediatrics 127, 994–1006. https://doi.org/10.1542/peds.2010-3500

Hesse, P.N., Schmitt, C., Klingenhoefer, S., Bremmer, F., 2017. Preattentive processing of numerical visual information. Front. Hum. Neurosci. 11, 1–14. https://doi.org/10.3389/fnhum.2017.00070

Hessl, D., Berry-Kravis, E., Cordeiro, L., Yuhas, J., Ornitz, E.M., Campbell, A., Chruscinski, E., Hervey, C., Long, J.M., Hagerman, R.J., 2009. Prepulse inhibition in fragile X syndrome: Feasibility, reliability, and implications for treatment. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. 150, 545–553. https://doi.org/10.1002/ajmg.b.30858

Hewitt, J.A., Brown, L.L., Murphy, S.J., Grieder, F., Silberberg, S.D., 2017. Accelerating biomedical discoveries through rigor and transparency. ILAR J. 58, 115–128. https://doi.org/10.1093/ilar/ilx011

Higgins, J.P.T., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. Stat. Med. 21, 1539–1558. https://doi.org/10.1002/sim.1186

Hodges, S.L., Nolan, S.O., Reynolds, C.D., Lugo, J.N., 2017. Spectral and temporal properties of calls reveal deficits in ultrasonic vocalizations of adult Fmr1 knockout mice. Behav. Brain Res. 332, 50–58. https://doi.org/10.1016/j.bbr.2017.05.052

Hodges, S.L., Reynolds, C.D., Nolan, S.O., Huebschman, J.L., Okoh, J.T., Binder, M.S., Lugo, J.N., 2019. A single early-life seizure results in long-term behavioral changes in the adult Fmr1 knockout mouse. Epilepsy Res. 157, 106193. https://doi.org/10.1016/j.eplepsyres.2019.106193

Hooijmans, C.R., Rovers, M.M., de Vries, R.B., Leenaars, M., Ritskes-Hoitinga, M., Langendam, M.W., 2014. SYRCLE's risk of bias tool for animal studies. BMC Med. Res. Methodol. 14. https://doi.org/10.1016/S0140-6736(02)09812-4

Hooper, A.W.M., Wong, H., Niibori, Y., Abdoli, R., Karumuthil-Melethil, S., Qiao, C., Danos, O., Bruder, J.T., Hampson, D.R., 2021. Gene therapy using an ortholog of human fragile X mental retardation protein partially rescues behavioral abnormalities and EEG activity. Mol. Ther. - Methods Clin. Dev. 22, 196–209. https://doi.org/10.1016/j.omtm.2021.06.013

Houtman, S.J., Lammertse, H.C.A., van Berkel, A.A., Balagura, G., Gardella, E., Ramautar, J.R., Reale, C., Møller, R.S., Zara, F., Striano, P., Misra-Isrie, M., van Haelst, M.M., Engelen, M., van Zuijen, T.L., Mansvelder, H.D., Verhage, M., Bruining, H., Linkenkaer-Hansen, K., 2021. STXBP1 Syndrome Is Characterized by Inhibition-Dominated Dynamics of Resting-State EEG. Front. Physiol. 12. https://doi.org/10.3389/fphys.2021.775172

Howland, J.G., Hannesson, D.K., Barnes, S.J., Phillips, A.G., 2007. Kindling of basolateral amygdala but not ventral hippocampus or perirhinal cortex disrupts sensorimotor gating in rats. Behav. Brain Res. 177, 30–36. https://doi.org/10.1016/j.bbr.2006.11.009

Hubel, D., Wiesel, T., 1968. Receptive Fields and Functional Architecture of Monkey Striate Cortex. J. Physiol. 195, 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455

Hubel, D.H., 1959. Single unit activity in striate cortex of unrestrained cats. J. Physiol. 147, 226–238. https://doi.org/10.1113/jphysiol.1959.sp006238

Huber, K.M., Klann, E., Costa-Mattioli, M., Zukin, R.S., 2015. Dysregulation of mammalian target of rapamycin signaling in mouse models of autism. J. Neurosci. 35, 13836. https://doi.org/10.1523/JNEUROSCI.2656-15.2015

Hunniford, V.T., Montroy, J., Fergusson, D.A., Avey, M.T., Wever, K.E., McCann, S.K., Foster, M., Fox, G., Lafreniere, M., Ghaly, M., Mannell, S., Godwinska, K., Gentles, A., Selim, S., MacNeil, J., Sikora, L., Sena, E.S., Page, M.J., Macleod, M., Moher, D., Lalu, M.M., 2021. Epidemiology and reporting characteristics of preclinical systematic reviews. PLoS Biol. 19, 1–17. https://doi.org/10.1371/journal.pbio.3001177

Inthout, J., Ioannidis, J.P.A., Borm, G.F., 2016. Obtaining evidence by a single well-powered trial or several modestly powered trials. Stat. Methods Med. Res. 25, 538–552. https://doi.org/10.1177/0962280212461098

Ioannidis, J.P.A., 2018. Why most published research findings are false. Get. to Good Res. Integr. Biomed. Sci. 2, 2–8. https://doi.org/10.1371/journal.pmed.0020124

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., Smith, J.D., Paeper, B., Nickerson, D.A., Dea, J., Dong, S., Gonzalez, L.E., Mandell, J.D., Mane, S.M., Murtha, M.T., Sullivan, C.A., Walker, M.F., Waqar, Z., Wei, L., Willsey, A.J., Yamrom, B., Lee, Y.H., Grabowska, E., Dalkic, E., Wang, Z., Marks, S., Andrews, P., Leotta, A., Kendall, J., Hakker, I., Rosenbaum, J., Ma, B., Rodgers, L., Troge, J., Narzisi, G., Yoon, S., Schatz, M.C., Ye, K., McCombie, W.R., Shendure, J., Eichler, E.E., State, M.W., Wigler, M., 2014. The contribution of de novo coding mutations to autism spectrum disorder. Nature 515, 216–221. https://doi.org/10.1038/nature13908

Ivanova, A., Zaidel, E., Salamon, N., Bookheimer, S., Uddin, L.Q., de Bode, S., 2017. Intrinsic functional organization of putative language networks in the brain following left cerebral hemispherectomy. Brain Struct. Funct. 222, 3795–3805. https://doi.org/10.1007/s00429-017-1434-y

Jacquemont, S., Berry-Kravis, E., Hagerman, R., Von Raison, F., Gasparini, F., Apostol, G., Ufer, M., Des Portes, V., Gomez-Mancilla, B., 2014. The challenges of clinical trials in fragile X syndrome. Psychopharmacology (Berl). 231, 1237–1250. https://doi.org/10.1007/s00213-013-3289-0

Jia, H., Yu, D., 2019. Attenuated long-range temporal correlations of electrocortical oscillations in patients with autism spectrum disorder. Dev. Cogn. Neurosci. 39, 100687. https://doi.org/10.1016/j.dcn.2019.100687

Jonak, C.R., Lovelace, J.W., Ethell, I.M., Razak, K.A., Binder, D.K., 2020. Multielectrode array analysis of EEG biomarkers in a mouse model of Fragile X Syndrome. Neurobiol. Dis. 138, 104794. https://doi.org/10.1016/j.nbd.2020.104794

Jones, S.R., 2016. When brain rhythms aren't 'rhythmic': implication for their mechanisms and meaning. Curr. Opin. Neurobiol. 40, 72–80. https://doi.org/10.1016/j.conb.2016.06.010

Jongs, N., Jagesar, R., van Haren, N.E.M., Penninx, B.W.J.H., Reus, L., Visser, P.J., van der Wee, N.J.A., Koning, I.M., Arango, C., Sommer, I.E.C., Eijkemans, M.J.C., Vorstman, J.A., Kas, M.J., 2020. A framework for assessing neuropsychiatric phenotypes by using smartphone-based location data. Transl. Psychiatry 10. https://doi.org/10.1038/s41398-020-00893-4

Juarez-Martinez, E.L., Sprengers, J.J., Cristian, G., Oranje, B., van Andel, D.M., Avramiea, A.E., Simpraga, S., Houtman, S.J., Hardstone, R., Gerver, C., Jan van der Wilt, G., Mansvelder, H.D., Eijkemans, M.J.C., Linkenkaer-Hansen, K., Bruining, H., 2021. Prediction of Behavioral Improvement Through Resting-State Electroencephalography and Clinical Severity in a Randomized Controlled Trial Testing Bumetanide in Autism Spectrum Disorder. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 1–11. https://doi.org/10.1016/j.bpsc.2021.08.009

Kadam, S.D., D'Ambrosio, R., Duveau, V., Roucard, C., Garcia-Cairasco, N., Ikeda, A., de Curtis, M., Galanopoulou, A.S., Kelly, K.M., 2017. Methodological standards and interpretation of video-electroencephalography in adult control rodents. A TASK1-WG1 report of the AES/ILAE Translational Task Force of the ILAE. Epilepsia 58, 10–27. https://doi.org/10.1111/epi.13903

Kafkafi, N., Agassi, J., Chesler, E.J., Crabbe, J.C., Crusio, W.E., Eilam, D., Gerlai, R., Golani, I., Gomez-Marin, A., Heller, R., Iraqi, F., Jaljuli, I., Karp, N.A., Morgan, H., Nicholson, G., Pfaff, D.W., Richter, S.H., Stark, P.B., Stiedl, O., Stodden, V., Tarantino, L.M., Tucci, V., Valdar, W., Williams, R.W., Würbel, H., Benjamini, Y., 2018. Reproducibility and replicability of rodent phenotyping in preclinical studies. Neurosci. Biobehav. Rev. 87, 218–232. https://doi.org/10.1016/j.neubiorev.2018.01.003

Kalueff, A. V., Stewart, A.M., Song, C., Berridge, K.C., Graybiel, A.M., Fentress, J.C., 2016. Neurobiology of rodent self-grooming and its value for translational neuroscience. Nat. Rev. Neurosci. 17, 45–59. https://doi.org/10.1038/nrn.2015.8

Karp, N.A., 2018. Reproducible preclinical research—Is embracing variability the answer? PLoS Biol. 16, 1–5. https://doi.org/10.1371/journal.pbio.2005413

Karp, N.A., Wilson, Z., Stalker, E., Mooney, L., Lazic, S.E., Zhang, B., Hardaker, E., 2020. A multi-batch design to deliver robust estimates of efficacy and reduce animal use – a syngeneic tumour case study. Sci. Rep. 10, 1–10. https://doi.org/10.1038/s41598-020-62509-7

Karpiel, I., Kurasz, Z., Kurasz, R., Duch, K., 2021. The influence of filters on EEG-ERP testing: Analysis of motor cortex in healthy subjects. Sensors 21, 1–18. https://doi.org/10.3390/s21227711

Kas, M.J., Penninx, B., Sommer, B., Serretti, A., Arango, C., Marston, H., 2019. A quantitative approach to neuropsychiatry: The why and the how. Neurosci. Biobehav. Rev. 97, 3–9. https://doi.org/10.1016/j.neubiorev.2017.12.008

Kaufmann, W.E., Kidd, S.A., Andrews, H.F., Budimirovic, D.B., Esler, A., Haas-Givler, B., Stackhouse, T., Riley, C., Peacock, G., Sherman, S.L., Brown, W.T., Berry-Kravis, E., 2017. Autism spectrum disorder in fragile X syndrome: Cooccurring conditions and current treatment. Pediatrics 139, S194–S206. https://doi.org/10.1542/peds.2016-1159F

Kazdoba, T.M., Leach, P.T., Silverman, J.L., Crawley, J.N., 2014. Modeling fragile X syndrome in the Fmr1 knockout mouse. Intractable rare Dis. Res. 3, 118–133. https://doi.org/10.5582/irdr.2014.01024

Keller, A.J., Martin, K.A.C., 2015. Local circuits for contrast normalization and adaptation investigated with two-photon imaging in cat primary visual cortex. J. Neurosci. 35, 10078–10087. https://doi.org/10.1523/JNEUROSCI.0906-15.2015

Kelley, D.J., Davidson, R.J., Elliott, J.L., Lahvis, G.P., Yin, J.C.P., Bhattacharyya, A., 2007. The cyclic AMP cascade is altered in the fragile X nervous system. PLoS One 2, 1–6. https://doi.org/10.1371/journal.pone.0000931

Kidd, S.A., Lachiewicz, A., Barbouth, D., Blitz, R.K., Delahunty, C., McBrien, D., Visootsak, J., Berry-Kravis, E., 2014. Fragile X syndrome: A review of associated medical problems. Pediatrics 134, 995–1005. https://doi.org/10.1542/peds.2013-4301

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F.W., Cuthill, I.C., Fry, D., Hutton, J., Altman, D.G., 2009. Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. PLoS One 4. https://doi.org/10.1371/journal.pone.0007824

Kimura, M., 2012. Visual mismatch negativity and unintentional temporal-context-based prediction in vision. Int. J. Psychophysiol. 83, 144–155. https://doi.org/10.1016/j.ijpsycho.2011.11.010

Kimura, M., Kondo, H., Ohira, H., Schröger, E., 2011. Unintentional temporal context-based prediction of emotional faces: An electrophysiological study. Cereb. Cortex 22, 1774–1785. https://doi.org/10.1093/cercor/bhr244

Kimura, M., Ohira, H., Schröger, E., 2010a. Localizing sensory and cognitive systems for pre-attentive visual deviance detection: An sLORETA analysis of the data of Kimura et al. (2009). Neurosci. Lett. 485, 198–203. https://doi.org/10.1016/j.neulet.2010.09.011

Kimura, M., Widmann, A., Schröger, E., 2010b. Human visual system automatically represents large-scale sequential regularities. Brain Res. 1317, 165–179. https://doi.org/10.1016/j.brainres.2009.12.076

Kissinger, S.T., Wu, Q., Quinn, C.J., Anderson, A.K., Pak, A., Chubykin, A.A., 2020. Visual Experience-Dependent Oscillations and Underlying Circuit Connectivity Changes Are Impaired in Fmr1 KO Mice. Cell Rep. 31. https://doi.org/10.1016/j.celrep.2020.03.050

Knoth, I.S., Vannasing, P., Major, P., Michaud, J.L., Lippé, S., 2014. Alterations of visual and auditory evoked potentials in fragile X syndrome. Int. J. Dev. Neurosci. 36, 90–97. https://doi.org/10.1016/j.ijdevneu.2014.05.003

Kogan, C.S., Boutet, I., Cornish, K., Graham, G.E., Berry-Kravis, E., Drouin, A., Milgram, N.W., 2009. A comparative neuropsychological test battery differentiates cognitive signatures of Fragile X and Down syndrome. J. Intellect. Disabil. Res. 53, 125–142. https://doi.org/10.1111/j.1365-2788.2008.01135.x

Kojouharova, P., File, D., Sulykos, I., Czigler, I., 2019. Visual mismatch negativity and stimulus-specific adaptation: the role of stimulus complexity. Exp. Brain Res. 237, 1179–1194. https://doi.org/10.1007/s00221-019-05494-2

Kokash, J., Alderson, E.M., Reinhard, S.M., Crawford, C.A., Binder, D.K., Ethell, I.M., Razak, K.A., 2019. Genetic reduction of MMP-9 in the Fmr1 KO mouse partially rescues prepulse inhibition of acoustic startle response. Brain Res. 1719, 24–29. https://doi.org/10.1016/j.brainres.2019.05.029

Korb, E., Herre, M., Zucker-Scharff, I., Gresack, J., Allis, C.D., Darnell, R.B., 2017. Excess Translation of Epigenetic Regulators Contributes to Fragile X Syndrome and Is Alleviated by Brd4 Inhibition. Cell 170, 1209-1223.e20. https://doi.org/10.1016/j.cell.2017.07.033

Kozono, N., Okamura, A., Honda, S., Matsumoto, M., Mihara, T., 2020. Gamma power abnormalities in a Fmr1-targeted transgenic rat model of fragile X syndrome. Sci. Rep. 10, 1–9. https://doi.org/10.1038/s41598-020-75893-x

Kreiner, G., 2015. Compensatory mechanisms in genetic models of neurodegeneration: Are the mice better than humans? Front. Cell. Neurosci. 9, 1–6. https://doi.org/10.3389/fncel.2015.00056

Kron, M., Howell, C.J., Adams, I.T., Ransbottom, M., Christian, D., Ogier, M., Katz, D.M., 2012. Brain activity mapping in Mecp2 mutant mice reveals functional deficits in forebrain circuits, including key nodes in the default mode network, that are reversed with ketamine treatment. J. Neurosci. 32, 13860–13872. https://doi.org/10.1523/JNEUROSCI.2159-12.2012

Krubitzer, L., 2009. In search of a unifying theory of complex brain evolution. Ann. New York Acad. Sci. 1156, 44–67. https://doi.org/10.1111/j.1749-6632.2009.04421.x

Krubitzer, L., 2007. The magnificent compromise: cortical field evolution in mammals. Neuron 56, 201–208. https://doi.org/10.1016/j.neuron.2007.10.002

Krubitzer, L.A., Seelke, A.M.H., 2012. Cortical evolution in mammals: The bane and beauty of phenotypic variability. Proc. Natl. Acad. Sci. U. S. A. 109, 10647–10654. https://doi.org/10.1073/pnas.1201891109

Kulinich, A.O., Reinhard, S.M., Rais, M., Lovelace, J.W., Scott, V., Binder, D.K., Razak, K.A., Ethell, I.M., 2020. Beneficial effects of sound exposure on auditory cortex development in a mouse model of Fragile X Syndrome. Neurobiol. Dis. 134, 104622. https://doi.org/10.1016/j.nbd.2019.104622

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest Package: Tests in Linear Mixed Effects Models. J. Stat. Softw. 82, 1–26. https://doi.org/10.18637/jss.v082.i13

Lacivita, E., Niso, M., Stama, M.L., Arzuaga, A., Altamura, C., Costa, L., Desaphy, J.F., Ragozzino, M.E., Ciranna, L., Leopoldo, M., 2020. Privileged scaffold-based design to identify a novel drug-like 5-HT7 receptor-preferring agonist to target Fragile X syndrome. Eur. J. Med. Chem. 199, 112395. https://doi.org/10.1016/j.ejmech.2020.112395

Landis, S.C., Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., Crystal, R.G., Darnell, R.B., Ferrante, R.J., Fillit, H., Finkelstein, R., Fisher, M., Gendelman, H.E., Golub, R.M., Goudreau,

J.L., Gross, R.A., Gubitz, A.K., Hesterlee, S.E., Howells, D.W., Huguenard, J., Kelner, K., Koroshetz, W., Krainc, D., Lazic, S.E., Levine, M.S., MacLeod, M.R., McCall, J.M., Iii, R.T.M., Narasimhan, K., Noble, L.J., Perrin, S., Porter, J.D., Steward, O., Unger, E., Utz, U., Silberberg, S.D., 2012. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490, 187–191. https://doi.org/10.1038/nature11556

Larson, J., Kim, D., Patel, R.C., Floreani, C., 2008. Olfactory discrimination learning in mice lacking the fragile X mental retardation protein. Neurobiol. Learn. Mem. 90, 90–102. https://doi.org/10.1016/j.nlm.2008.01.002

Lee, E., Lee, J., Kim, E., 2017. Excitation / Inhibition Imbalance in Animal Models of Autism Spectrum Disorders. Biol. Psychiatry 81, 838–847. https://doi.org/10.1016/j.biopsych.2016.05.011

Lee, F.H.F., Lai, T.K.Y., Su, P., Liu, F., 2019. Altered cortical Cytoarchitecture in the Fmr1 knockout mouse. Mol. Brain 12, 1–12. https://doi.org/10.1186/s13041-019-0478-8

Lee, J., Chung, C., Ha, S., Lee, D., Kim, D.Y., Kim, H., Kim, E., 2015. Shank3-mutant mice lacking exon 9 show altered excitation/inhibition balance, enhanced rearing, and spatial memory deficit. Front. Cell. Neurosci. 9, 1–14. https://doi.org/10.3389/fncel.2015.00094

Lee, M., Balla, A., Sershen, H., Sehatpour, P., Lakatos, P., Javitt, D.C., 2018. Rodent Mismatch Negativity/theta Neuro-Oscillatory Response as a Translational Neurophysiological Biomarker for N-Methyl-D-Aspartate Receptor-Based New Treatment Development in Schizophrenia. Neuropsychopharmacology 43, 571–582. https://doi.org/10.1038/npp.2017.176

Leigh, M.J.S., Nguyen, D. V, Mu, Y., Winarni, T.I., Schneider, A., Chechi, T., Polussa, J., Doucet, P., Tassone, F., Rivera, S.M., Hessl, D., Hagerman, R.J., 2013. A randomized double-blind, placebo-controlled trial of minocycline in children and adolescents with fragile X syndrome. J. Dev. Behav. Pediatr. 34, 147–155.

Leiser, S.C., Dunlop, J., Bowlby, M.R., Devilbiss, D.M., 2011. Aligning strategies for using EEG as a surrogate biomarker: A review of preclinical and clinical research. Biochem. Pharmacol. 81, 1408–1421. https://doi.org/10.1016/j.bcp.2010.10.002

Li, J., Jiang, R.Y., Arendt, K.L., Hsu, Y.T., Zhai, S.R., Chen, L., 2020. Defective memory engram reactivation underlies impaired fear memory recall in fragile x syndrome. Elife 9, 1–20. https://doi.org/10.7554/eLife.61882

Liao, W., Chen, S., Huang, J., Xiang, H., Wei, J., Pan, Z., Zhang, S., Ye, Z., Cai, H., Pan, Y., 2018. α-Asarone ameliorated learning and memory ability in fragile x syndrome model mice via down-regulating p-ERK1/2 expression. Int. J. Clin. Exp. Med. 11, 1–9.

Linkenkaer-Hansen, K., Monto, S., Rytsälä, H., Suominen, K., Isometsä, E., Kähkönen, S., 2005. Breakdown of long-range temporal correlations in theta oscillations in patients with major depressive disorder. J. Neurosci. 25, 10131–10137. https://doi.org/10.1523/JNEUROSCI.3244-05.2005

Linkenkaer-Hansen, K., Smit, D.J.A., Barkil, A., Van Beijsterveldt, T.E.M., Brussaard, A.B., Boomsma, D.I., Van Ooyen, A., De Geus, E.J.C., 2007. Genetic contributions to long-range temporal correlations in ongoing oscillations. J. Neurosci. 27, 13882–13889. https://doi.org/10.1523/JNEUROSCI.3083-07.2007

Liu, Z.-H., Smith, C.B., 2009. Dissociation of social and nonsocial anxiety in a mouse model of fragile X syndrome. Neurosci. Lett. 454, 62–66. https://doi.org/10.1016/j.neulet.2009.02.066

Livingston, L.A., Happé, F., 2017. Conceptualising compensation in neurodevelopmental disorders: Reflections from autism spectrum disorder. Neurosci. Biobehav. Rev. 80, 729–742. https://doi.org/10.1016/j.neubiorev.2017.06.005

Lopez, L., Brusa, A., Fadda, A., Loizzo, S., Martinangeli, A., Sannita, W.G., Loizzo, A., 2002. Modulation of flash stimulation intensity and frequency: Effects on visual evoked potentials and oscillatory

potentials recorded in awake, freely moving mice. Behav. Brain Res. 131, 105–114. https://doi.org/10.1016/S0166-4328(01)00351-5

Louhivuori, V., Vicario, A., Uutela, M., Rantamäki, T., Louhivuori, L.M., Castrén, E., Tongiorgi, E., Åkerman, K.E., Castrén, M.L., 2011. BDNF and TrkB in neuronal differentiation of Fmr1-knockout mouse. Neurobiol. Dis. 41, 469–480. https://doi.org/10.1016/j.nbd.2010.10.018

Lovelace, J.W., Ethell, I.M., Binder, D.K., Razak, K.A., 2020. Minocycline Treatment Reverses Sound Evoked EEG Abnormalities in a Mouse Model of Fragile X Syndrome. Front. Neurosci. 14, 1–16. https://doi.org/10.3389/fnins.2020.00771

Lovelace, J.W., Ethell, I.M., Binder, D.K., Razak, K.A., 2018. Translation-relevant EEG phenotypes in a mouse model of Fragile X Syndrome. Neurobiol. Dis. 115, 39–48. https://doi.org/10.1016/j.nbd.2018.03.012

Lovelace, J.W., Wen, T.H., Reinhard, S., Hsu, M.S., Sidhu, H., Ethell, I.M., Binder, D.K., Razak, K.A., 2016. Matrix metalloproteinase-deletion rescues auditory evoked potential habituation deficit in a mouse model of Fragile X Syndrome. Neurobiol. Dis. 89, 126–135. https://doi.org/10.1016/j.nbd.2016.02.002

Luhmann, H.J., Sinning, A., Yang, J.-W., Reyes-Puerta, V., Stüttgen, M.C., Kirischuk, S., Kilb, W., 2016. Spontaneous Neuronal Activity in Developing Neocortical Networks: From Single Cells to Large-Scale Interactions. Front. Neural Circuits 10, 1–14. https://doi.org/10.3389/fncir.2016.00040

Luke, S.G., 2017. Evaluating significance in linear mixed-effects models in R. Behav. Res. Methods 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

MacLeod, L.S., Kogan, C.S., Collin, C.A., Berry-Kravis, E., Messier, C., Gandhi, R., 2010. A comparative study of the performance of individuals with fragile X syndrome and Fmr1 knockout mice on Hebb-Williams mazes. Genes, Brain Behav. 9, 53–64. https://doi.org/10.1111/j.1601-183X.2009.00534.x

Maheshwari, A., 2020. Rodent EEG: Expanding the Spectrum of Analysis. Epilepsy Curr. 20, 149–153. https://doi.org/10.1177/1535759720921377

Mancuso, C.E., Tanzi, M.G., Gabay, M., 2004. Paradoxical reactions to benzodiazepines: Literature review and treatment options. Pharmacotherapy 24, 1177–1185. https://doi.org/10.1592/phco.24.13.1177.38089

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Matuoka, T., Yabe, H., Shinozaki, N., Sato, Y., Hiruma, T., Ren, A., Hara, E., Kaneko, S., 2006. The Development of Memory Trace Depending on the Number of the Standard Stimuli. Clin. EEG Neurosci. 37, 223–229. https://doi.org/10.1177/155005940603700312

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., Bates, D., 2017. Balancing Type I error and power in linear mixed models. J. Mem. Lang. 94, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

May, P., Tiitinen, H., Ilmoniemi, R.J., Nyman, G., Taylor, J.G., Näätänen, R., 1999. Frequency change detection in human auditory cortex. J. Comput. Neurosci. 6, 99–120. https://doi.org/10.1023/A:1008896417606

McGraw, C.M., Ward, C.S., Samaco, R.C., 2017. Genetic rodent models of brain disorders: Perspectives on experimental approaches and therapeutic strategies. Am. J. Med. Genet. Part C Semin. Med. Genet. 175, 368–379. https://doi.org/10.1002/ajmg.c.31570

McNaughton, C.H., Moon, J., Strawderman, M.S., Maclean, K.N., Evans, J., Strupp, B.J., 2008. Evidence for Social Anxiety and Impaired Social Cognition in a Mouse Model of Fragile X Syndrome. Behav. Neurosci. 122, 293–300. https://doi.org/10.1037/0735-7044.122.2.293

Melancia, F., Trezza, V., 2018. Modelling fragile X syndrome in the laboratory setting: A behavioral perspective. Behav. Brain Res. 350, 149–163. https://doi.org/10.1016/j.bbr.2018.04.042

Melnik, A., Legkov, P., Izdebski, K., Kärcher, S.M., Hairston, W.D., Ferris, D.P., König, P., 2017. Systems, subjects, sessions: To what extent do these factors influence EEG data? Front. Hum. Neurosci. 11, 1–20. https://doi.org/10.3389/fnhum.2017.00150

Michalon, A., Sidorov, M., Ballard, T.M., Ozmen, L., Spooren, W., Wettstein, J.G., Jaeschke, G., Bear, M.F., Lindemann, L., 2012. Chronic Pharmacological mGlu5 Inhibition Corrects Fragile X in Adult Mice. Neuron 74, 49–56. https://doi.org/10.1016/j.neuron.2012.03.009

Mientjes, E.J., Nieuwenhuizen, I., Kirkpatrick, L., Zu, T., Hoogeveen-Westerveld, M., Severijnen, L., Rifé, M., Willemsen, R., Nelson, D.L., Oostra, B.A., 2006. The generation of a conditional Fmr1 knock out mouse model to study Fmrp function in vivo. Neurobiol. Dis. 21, 549–555. https://doi.org/10.1016/j.nbd.2005.08.019

Mines, M.A., Yuskaitis, C.J., King, M.K., Beurel, E., Jope, R.S., 2010. GSK3 influences social preference and anxiety-related behaviors during social interaction in a mouse model of fragile X syndrome and autism. PLoS One 5. https://doi.org/10.1371/journal.pone.0009706

Molenhuis, R.T., Bruining, H., Brandt, M.J.V., Van Soldt, P.E., Abu-Toamih Atamni, H.J., Burbach, J.P.H., Iraqi, F.A., Mott, R.F., Kas, M.J.H., 2018. Modeling the quantitative nature of neurodevelopmental disorders using Collaborative Cross mice. Mol. Autism 9, 1–11. https://doi.org/10.1186/s13229-018-0252-2

Montez, T., Poil, S.S., Jones, B.F., Manshanden, I., Verbunt, J.P.A., Van Dijk, B.W., Brussaard, A.B., Van Ooyen, A., Stam, C.J., Scheltens, P., Linkenkaer-Hansen, K., 2009. Altered temporal correlations in parietal alpha and prefrontal theta oscillations in early-stage Alzheimer disease. Proc. Natl. Acad. Sci. U. S. A. 106, 1614–1619. https://doi.org/10.1073/pnas.0811699106

Montijn, J.S., Olcese, U., Pennartz, C.M.A., 2016. Visual stimulus detection correlates with the consistency of temporal sequences within stereotyped events of V1 neuronal population activity. J. Neurosci. 36, 8624–8640. https://doi.org/10.1523/JNEUROSCI.0853-16.2016

Monto, S., Vanhatalo, S., Holmes, M.D., Palva, J.M., 2007. Epileptogenic neocortical networks are revealed by abnormal temporal dynamics in seizure-free subdural EEG. Cereb. Cortex 17, 1386–1393. https://doi.org/10.1093/cercor/bhl049

Morcom, A.M., Johnson, W., 2015. Neural reorganization and compensation in ageing. J. Cogn. Neurosci. 27, 1275–1285. https://doi.org/10.1162/jocn_a_00783

Moy, S.S., Nadler, J.J., Young, N.B., Nonneman, R.J., Grossman, A.W., Murphy, D.L., D'Ercole, A.J., Crawley, J.N., Magnuson, T.R., Lauder, J.M., 2009. Social approach in genetically engineered mouse lines relevant to autism. Genes. Brain. Behav. 8, 129–142. https://doi.org/10.1111/j.1601-183X.2008.00452.x

Moyer, J.T., Gnatkovsky, V., Ono, T., Otáhal, J., Wagenaar, J., Stacey, W.C., Noebels, J., Ikeda, A., Staley, K., de Curtis, M., Litt, B., Galanopoulou, A.S., 2017. Standards for data acquisition and software-based analysis of in vivo electroencephalography recordings from animals. A TASK1-WG5 report of the AES/ILAE Translational Task Force of the ILAE. Epilepsia 58, 53–67. https://doi.org/10.1111/epi.13909

Muthukumaraswamy, S.D., 2013. High-frequency brain activity and muscle artifacts in MEG/EEG: A review and recommendations. Front. Hum. Neurosci. 7, 1–11. https://doi.org/10.3389/fnhum.2013.00138

Näätänen, R., Sussman, E.S., Salisbury, D., Shafer, V.L., 2014. Mismatch negativity (MMN) as an index of cognitive dysfunction. Brain Topogr. 27, 451–466. https://doi.org/10.1007/s10548-014-0374-6

Nakagawa, S., Schielzeth, H., 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol. Evol. 4, 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Napoli, E., Ross-Inta, C., Song, G., Wong, S., Hagerman, R., Gane, L.W., Smilowitz, J.T., Tassone, F., Giulivi, C., 2016. Premutation in the Fragile X Mental Retardation 1 (FMR1) gene affects maternal Zn-

milk and perinatal brain bioenergetics and scaffolding. Front. Neurosci. 10, 1–26. https://doi.org/10.3389/fnins.2016.00159

Naviaux, J.C., Wang, L., Li, K., Bright, A.T., Alaynick, W.A., Williams, K.R., Powell, S.B., Naviaux, R.K., 2015. Antipurinergic therapy corrects the autism-like features in the Fragile X (Fmr1 knockout) mouse model. Mol. Autism 6, 1. https://doi.org/10.1186/2040-2392-6-1

Nelson, S.B., Valakh, V., 2015. Excitatory/Inhibitory Balance and Circuit Homeostasis in Autism Spectrum Disorders. Neuron 87, 684–698. https://doi.org/10.1016/j.neuron.2015.07.033

Nikulin, V. V., Jönsson, E.G., Brismar, T., 2012. Attenuation of long-range temporal correlations in the amplitude dynamics of alpha and beta neuronal oscillations in patients with schizophrenia. Neuroimage 61, 162–169. https://doi.org/10.1016/j.neuroimage.2012.03.008

Niu, M., Han, Y., Dy, A.B.C., Du, J., Jin, H., Qin, J., Zhang, J., Li, Q., Hagerman, R.J., 2017. Autism Symptoms in Fragile X Syndrome. J. Child Neurol. 32, 903–909. https://doi.org/10.1177/0883073817712875

Nolan, S.O., Reynolds, C.D., Smith, G.D., Holley, A.J., Escobar, B., Chandler, M.A., Volquardsen, M., Jefferson, T., Pandian, A., Smith, T., Huebschman, J., Lugo, J.N., 2017. Deletion of Fmr1 results in sex-specific changes in behavior. Brain Behav. 7, 1–13. https://doi.org/10.1002/brb3.800

Ollikainen, J.O., Vauhkonen, M., Karjalainen, P.A., Kaipio, J.P., 2000. Effects of electrode properties on EEG measurements and a related inverse problem. Med. Eng. Phys. 22, 535–545. https://doi.org/10.1016/S1350-4533(00)00070-9

Olmos-Serrano, J.L., Corbin, J.G., Burns, M.P., 2011. The GABA A receptor agonist THIP ameliorates specific behavioral deficits in the mouse model of fragile X syndrome. Dev. Neurosci. 33, 395–403. https://doi.org/10.1159/000332884

Olmos-Serrano, J.L., Paluszkiewicz, S.M., Martin, B.S., Kaufmann, W.E., Corbin, J.G., Huntsman, M.M., 2010. Defective GABAergic neurotransmission and pharmacological rescue of neuronal hyperexcitability in the amygdala in a mouse model of fragile X syndrome. J. Neurosci. 30, 9929–9938. https://doi.org/10.1523/JNEUROSCI.1714-10.2010

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., 2011. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. 2011. https://doi.org/10.1155/2011/156869

Orefice, L.L.L., Zimmerman, A.L.L., Chirila, A.M.M., Sleboda, S.J.J., Head, J.P.P., Ginty, D.D.D., 2016. Peripheral Mechanosensory Neuron Dysfunction Underlies Tactile and Behavioral Deficits in Mouse Models of ASDs. Cell 166, 299–314. https://doi.org/10.1016/j.cell.2016.05.033

Ouzzani, M., Hammady, H., Fedorowicz, Z., Elmagarmid, A., 2016. Rayyan-a web and mobile app for systematic reviews. Syst. Rev. 5, 1–10. https://doi.org/10.1186/s13643-016-0384-4

Pacey, L.K.K., Doss, L., Cifelli, C., van der Kooy, D., Heximer, S.P., Hampson, D.R., 2011. Genetic deletion of regulator of G-protein signaling 4 (RGS4) rescues a subset of fragile X related phenotypes in the FMR1 knockout mouse. Mol. Cell. Neurosci. 46, 563–572. https://doi.org/10.1016/j.mcn.2010.12.005

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D., 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ 372. https://doi.org/10.1136/bmj.n71

Pak, A., Kissinger, S.T., Chubykin, A.A., 2021. Impaired Adaptation and Laminar Processing of the Oddball Paradigm in the Primary Visual Cortex of Fmr1 KO Mouse. Front. Cell. Neurosci. 15, 1–15. https://doi.org/10.3389/fncel.2021.668230

Paluszkiewicz, S.M., Martin, B.S., Huntsman, M.M., 2011. Fragile X syndrome: The GABAergic system and circuit dysfunction. Dev. Neurosci. 33, 349–364. https://doi.org/10.1159/000329420

Papazoglou, A., Lundt, A., Wormuth, C., Ehninger, D., Henseler, C., Soós, J., Broich, K., Weiergräber, M., 2016. Non-restraining EEG radiotelemetry: Epidural and deep intracerebral stereotaxic EEG electrode placement. J. Vis. Exp. 2016, 1–16. https://doi.org/10.3791/54216

Paribello, C., Tao, L., Folino, A., Berry-Kravis, E., Tranfaglia, M., Ethell, I.M., Ethell, D.W., 2010. Open-label add-on treatment trial of minocycline in fragile X syndrome. BMC Neurol. 10, 1–9. https://doi.org/10.1186/1471-2377-10-91

Park, J., van den Berg, B., Chiang, C., Woldorff, M.G., Brannon, E.M., 2018. Developmental trajectory of neural specialization for letter and number visual processing. Dev. Sci. 21, 1–14. https://doi.org/10.1111/desc.12578

Paylor, R., Yuva-Paylor, L.A., Nelson, D.L., Spencer, C.M., 2008. Reversal of sensorimotor gating abnormalities in Fmr1 knockout mice carrying a human Fmr1 transgene. Behav. Neurosci. 122, 1371–1377. https://doi.org/10.1037/a0013047

Pazo-Alvarez, P., Cadaveira, F., Amenedo, E., 2003. MMN in the visual modality: A review. Biol. Psychol. 63, 199–236. https://doi.org/10.1016/S0301-0511(03)00049-8

Peleh, T., Ike, K.G.O., Frentz, I., Buwalda, B., de Boer, S.F., Hengerer, B., Kas, M.J.H., 2020. Cross-site Reproducibility of Social Deficits in Group-housed BTBR Mice Using Automated Longitudinal Behavioural Monitoring. Neuroscience 445, 95–108. https://doi.org/10.1016/j.neuroscience.2020.04.045

Peleh, T., Ike, K.G.O., Wams, E.J., Lebois, E.P., Hengerer, B., 2019. The reverse translation of a quantitative neuropsychiatric framework into preclinical studies: Focus on social interaction and behavior. Neurosci. Biobehav. Rev. 97, 96–111. https://doi.org/10.1016/j.neubiorev.2018.07.018

Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M.T., Baker, M., Browne, W.J., Clark, A., Cuthill, I.C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S.T., Howells, D.W., Karp, N.A., Lazic, S.E., Lidster, K., MacCallum, C.J., Macleod, M., Pearl, E.J., Petersen, O.H., Rawle, F., Reynolds, P., Rooney, K., Sena, E.S., Silberberg, S.D., Steckler, T., Würbel, H., 2020. The arrive guidelines 2.0: Updated guidelines for reporting animal research. PLoS Biol. 18, 1–12. https://doi.org/10.1371/journal.pbio.3000410

Perenboom, T., Schenke, M., Ferrari, M., Terwindt, G., van den Maagdenberg, A., Tolner, E., 2020. Responsivity to light in familial hemiplegic migraine type 1 mutant mice reveals frequency-dependent enhancement of visual network excitability. Eur. J. Neurosci. 53, 1672–1686. https://doi.org/10.1111/ejn.15041

Pietropaolo, S., Guilleminot, A., Martin, B., D'Amato, F.R., Crusio, W.E., 2011. Genetic-background modulation of core and variable autistic-like symptoms in Fmr1 knock-out mice. PLoS One 6, 1–11. https://doi.org/10.1371/journal.pone.0017073

Pirbhoy, P.S., Rais, M., Lovelace, J.W., Woodard, W., Razak, K.A., Binder, D.K., Ethell, I.M., 2020. Acute pharmacological inhibition of matrix metalloproteinase-9 activity during development restores perineuronal net formation and normalizes auditory processing in Fmr1 KO mice. J. Neurochem. https://doi.org/10.1111/jnc.15037

Portis, S., Giunta, B., Obregon, D., Tan, J., 2012. The role of glycogen synthase kinase-3 signaling in neurodevelopment and fragile X syndrome. Int. J. Physiol. Pathophysiol. Pharmacol. 4, 140–148.

Prinz, F., Schlange, T., Asadullah, K., 2011. Believe it or not: How much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. 10, 712–713. https://doi.org/10.1038/nrd3439-c1

Prut, L., Belzung, C., 2003. The open field as a paradigm to measure the effects of drugs on anxiety-like behaviors: A review. Eur. J. Pharmacol. 463, 3–33. https://doi.org/10.1016/S0014-2999(03)01272-X

Puścian, A., Łęski, S., Kasprowicz, G., Winiarski, M., Borowska, J., Nikolaev, T., Boguszewski, P.M., Lipp, H.P., Knapska, E., 2016. Eco-HAB as a fully automated and ecologically relevant assessment of social impairments in mouse models of autism. Elife 5, 1–22. https://doi.org/10.7554/eLife.19532

Pyronneau, A., He, Q., Hwang, J.Y., Porch, M., Contractor, A., Zukin, R.S., 2017. Aberrant Rac1-cofilin signaling mediates defects in dendritic spines, synaptic function, and sensory perception in fragile X syndrome. Sci. Signal. 10, 1–16. https://doi.org/10.1126/scisignal.aan0852

Qin, M., Zeidler, Z., Moulton, K., Krych, L., Xia, Z., Smith, C.B., 2015. Endocannabinoid-mediated improvement on a test of aversive memory in a mouse model of fragile X syndrome. Behav. Brain Res. 291, 164–171. https://doi.org/10.1016/j.bbr.2015.05.003

Rais, M., Binder, D.K., Razak, K.A., Ethell, I.M., 2018. Sensory Processing Phenotypes in Fragile X Syndrome. ASN Neuro 10, 1759091418801092. https://doi.org/10.1177/1759091418801092

Ramírez-López, A., Pastor, A., de la Torre, R., La Porta, C., Ozaita, A., Cabañero, D., Maldonado, R., 2021. Role of the endocannabinoid system in a mouse model of Fragile X undergoing neuropathic pain. Eur. J. Pain 25, 1316–1328. https://doi.org/10.1002/ejp.1753

Ramsteijn, A.S., Van de Wijer, L., Rando, J., van Luijk, J., Homberg, J.R., Olivier, J.D.A., 2020. Perinatal selective serotonin reuptake inhibitor exposure and behavioral outcomes: A systematic review and meta-analyses of animal studies. Neurosci. Biobehav. Rev. 114, 53–69. https://doi.org/10.1016/j.neubiorev.2020.04.010

Reeb-Sutherland, B.C., Fox, N.A., 2015. Eyeblink Conditioning: A Non-invasive Biomarker for Neurodevelopmental Disorders. J. Autism Dev. Disord. 45, 376–394. https://doi.org/10.1007/s10803-013-1905-9

Restivo, L., Ferrari, F., Passino, E., Sgobio, C., Bock, J., Oostra, B.A., Bagni, C., Ammassari-Teule, M., 2005. Enriched environment promotes behavioral and morphological recovery in a mouse model for the fragile X syndrome. Proc. Natl. Acad. Sci. U. S. A. 102, 11557–11562. https://doi.org/10.1073/pnas.0504984102

Richter, H.S., 2017. Systematic heterogenization for better reproducibility in animal experimentation. Lab Anim. (NY). 46, 343–349. https://doi.org/10.1038/laban.1330

Richter, S.H., Garner, J.P., Auer, C., Kunert, J., Würbel, H., 2010. Systematic variation improves reproducibility of animal experiments. Nat. Methods 7, 167–168. https://doi.org/10.1038/nmeth0310-167.

Richter, S.H., Garner, J.P., Würbel, H., 2009. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nat. Methods 6, 257–261. https://doi.org/10.1038/nmeth.1312.

Richter, S.H., Garner, J.P., Zipser, B., Lewejohann, L., Sachser, N., Touma, C., Schindler, B., Chourbaji, S., Brandwein, C., Gass, P., van Stipdonk, N., van der Harst, J., Spruijt, B., Võikar, V., Wolfer, D.P., Würbel, H., 2011. Effect of population heterogenization on the reproducibility of mouse behavior: A multi-laboratory study. PLoS One 6. https://doi.org/10.1371/journal.pone.0016461

Ridder, W.H., Nusinowitz, S., 2006. The visual evoked potential in the mouse - Origins and response characteristics. Vision Res. 46, 902–913. https://doi.org/10.1016/j.visres.2005.09.006

Rigoulot, S., Knoth, I.S., Lafontaine, M.P., Vannasing, P., Major, P., Jacquemont, S., Michaud, J.L., Jerbi, K., Lippé, S., 2017. Altered visual repetition suppression in Fragile X Syndrome: New evidence from ERPs and oscillatory activity. Int. J. Dev. Neurosci. 59, 52–59. https://doi.org/10.1016/j.ijdevneu.2017.03.008

Rodrigues, J., Studer, E., Streuber, S., Meyer, N., Sandi, C., 2020. Locomotion in virtual environments predicts cardiovascular responsiveness to subsequent stressful challenges. Nat. Commun. 11, 1–11. https://doi.org/10.1038/s41467-020-19736-3

Rotschafer, S.E., Razak, K.A., 2014. Auditory processing in fragile x syndrome. Front. Cell. Neurosci. 8, 19. https://doi.org/10.3389/fncel.2014.00019

Rubenstein, J.L.R., Merzenich, M.M., 2003. Model of autism : increased ratio of excitation / inhibition in key neural systems. Genes, Brain Behav. 2, 255–267. https://doi.org/10.1046/j.1601-183X.2003.00037.x

Sabanov, V., Braat, S., D'Andrea, L., Willemsen, R., Zeidler, S., Rooms, L., Bagni, C., Kooy, R.F., Balschun, D., 2017. Impaired GABAergic inhibition in the hippocampus of Fmr1 knockout mice. Neuropharmacology 116, 71–81. https://doi.org/10.1016/j.neuropharm.2016.12.010

Sabri, M., Campbell, K.B., 2001. Effects of sequential and temporal probability of deviant occurrence on mismatch negativity. Cogn. Brain Res. 12, 171–180. https://doi.org/10.1016/S0926-6410(01)00026-X

Saleem, A.B., Lien, A.D., Krumin, M., Haider, B., Rosón, M.R., Ayaz, A., Reinhold, K., Busse, L., Carandini, M., Harris, K.D., Carandini, M., 2017. Subcortical Source and Modulation of the Narrowband Gamma Oscillation in Mouse Visual Cortex. Neuron 93, 315–322. https://doi.org/10.1016/j.neuron.2016.12.028

Sambeth, A., Maes, J.H.R., Van Luijtelaar, G., Molenkamp, I.B.S., Jongsma, M.L.A., Van Rijn, C.M., 2003. Auditory event-related potentials in humans and rats: Effects of task manipulation. Psychophysiology 40, 60–68. https://doi.org/10.1111/1469-8986.00007

Sanchez-Vives, M. V., Nowak, L.G., McCormick, D.A., 2000. Cellular mechanisms of long-lasting adaptation in visual cortical neurons in vitro. J. Neurosci. 20, 4286–4299. https://doi.org/10.1523/jneurosci.20-11-04286.2000

Saré, R.M., Figueroa, C., Lemons, A., Loutaev, I., Beebe Smith, C., 2019. Comparative Behavioral Phenotypes of Fmr1 KO, Fxr2 Het, and Fmr1 KO/Fxr2 Het Mice. Brain Sci. 9. https://doi.org/10.3390/brainsci9010013

Saré, R.M., Levine, M., Smith, C.B., 2016. Behavioral Phenotype of Fmr1 Knock-Out Mice during Active Phase in an Altered Light/Dark Cycle. eNeuro 3. https://doi.org/10.1523/ENEURO.0035-16.2016

Saré, R.M., Song, A., Loutaev, I., Cook, A., Maita, I., Lemons, A., Sheeler, C., Smith, C.B., 2018. Negative effects of chronic rapamycin treatment on behavior in a mouse model of fragile X syndrome. Front. Mol. Neurosci. 10, 1–11. https://doi.org/10.3389/fnmol.2017.00452

Sassenhagen, J., Draschkow, D., 2019. Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. Psychophysiology 56, 1–8. https://doi.org/10.1111/psyp.13335

Sato, Y., Yabe, H., Hiruma, T., Sutoh, T., Shinozaki, N., Nashida, T., Kaneko, S., 2000. The effect of deviant stimulus probability on the human mismatch process. Neuroreport 11, 3703–3708. https://doi.org/10.1097/00001756-200011270-00023

Scanlon, J.E.M., Redman, E.X., Kuziek, J.W.P., Mathewson, K.E., 2020. A ride in the park: Cycling in different outdoor environments modulates the auditory evoked potentials. Int. J. Psychophysiol. 151, 59–69. https://doi.org/10.1016/j.ijpsycho.2020.02.016

Scanlon, J.E.M., Townsend, K.A., Cormier, D.L., Kuziek, J.W.P., Mathewson, K.E., 2019. Taking off the training wheels: Measuring auditory P3 during outdoor cycling using an active wet EEG system. Brain Res. 1716, 50–61. https://doi.org/10.1016/j.brainres.2017.12.010

Schaefer, T.L., Ashworth, A.A., Tiwari, D., Tomasek, M.P., Parkins, E. V., White, A.R., Snider, A., Davenport, M.H., Grainger, L.M., Becker, R.A., Robinson, C.K., Mukherjee, R., Williams, M.T., Gibson, J.R., Huber, K.M., Gross, C., Erickson, C.A., 2021. GABAA Alpha 2,3 Modulation Improves Select Phenotypes in a Mouse Model of Fragile X Syndrome. Front. Psychiatry 12, 1–15. https://doi.org/10.3389/fpsyt.2021.678090

Schilit Nitenson, A., Stackpole, E.E., Truszkowski, T.L.S., Midroit, M., Fallon, J.R., Bath, K.G., 2015. Fragile X mental retardation protein regulates olfactory sensitivity but not odorant discrimination. Chem. Senses 40, 345–350. https://doi.org/10.1093/chemse/bjv019

Schmeisser, M.J., Ey, E., Wegener, S., Bockmann, J., Stempel, A.V., Kuebler, A., Janssen, A.L., Udvardi, P.T., Shiban, E., Spilker, C., Balschun, D., Skryabin, B. V., Dieck, S.T., Smalla, K.H., Montag, D., Leblond, C.S., Faure, P., Torquet, N., Le Sourd, A.M., Toro, R., Grabrucker, A.M., Shoichet, S.A., Schmitz, D., Kreutz, M.R., Bourgeron, T., Gundelfinger, E.D., Boeckers, T.M., 2012. Autistic-like behaviours and hyperactivity in mice lacking ProSAP1/Shank2. Nature 486, 256–260. https://doi.org/10.1038/nature11015

Schmitt, L.M., Shaffer, R.C., Hessl, D., Erickson, C., 2019. Executive function in fragile X syndrome: A systematic review. Brain Sci. 9. https://doi.org/10.3390/brainsci9010015

Sculthorpe, L.D., Ouellet, D.R., Campbell, K.B., 2009. MMN elicitation during natural sleep to violations of an auditory pattern. Brain Res. 1290, 52–62. https://doi.org/10.1016/j.brainres.2009.06.013

Semple, B.D., Blomgren, K., Gimlin, K., Ferriero, D.M., Noble-Haeusslein, L.J., 2013. Brain development in rodents and humans: Identifying benchmarks of maturation and vulnerability to injury across species. Prog. Neurobiol. 106–107, 1–16. https://doi.org/10.1016/j.pneurobio.2013.04.001

Sena, E.S., Currie, G.L., McCann, S.K., Macleod, M.R., Howells, D.W., 2014. Systematic reviews and meta-analysis of preclinical studies: Why perform them and how to appraise them critically. J. Cereb. Blood Flow Metab. 34, 737–742. https://doi.org/10.1038/jcbfm.2014.28

Sengupta, P., 2013. The laboratory rat: Relating its age with human's. Int. J. Prev. Med. 4, 624–630.

Senzai, Y., Fernandez-Ruiz, A., Buzsáki, G., 2019. Layer-Specific Physiological Features and Interlaminar Interactions in the Primary Visual Cortex of the Mouse. Neuron 101, 500-513.e5. https://doi.org/10.1016/j.neuron.2018.12.009

Sherry, C.E., Pollard, J.Z., Tritz, D., Carr, B.K., Pierce, A., Vassar, M., 2020. Assessment of transparent and reproducible research practices in the psychiatry literature. Gen. Psychiatry 33. https://doi.org/10.1136/gpsych-2019-100149

Shili, I., Hamdi, Y., Marouani, A., Ben Lasfar, Z., Ghrairi, T., Lefranc, B., Leprince, J., Vaudry, D., Olfa, M.K., 2021. Long-term protective effect of PACAP in a fetal alcohol syndrome (FAS) model. Peptides 146. https://doi.org/10.1016/j.peptides.2021.170630

Shouse, J.N., Rowe, S. V., Mast, B.T., 2013. Depression and Cognitive Functioning as Predictors of Social Network Size. Clin. Gerontol. 36, 147–161. https://doi.org/10.1080/07317115.2012.749320

Siegel, S.J., Connolly, P., Liang, Y., Lenox, R.H., Gur, R.E., Bilker, W.B., Kanes, S.J., Turetsky, B.I., 2003. Effects of strain, novelty, and NMDA blockade on auditory-evoked potentials in mice. Neuropsychopharmacology 28, 675–682. https://doi.org/10.1038/sj.npp.1300087

Sil, A., Bespalov, A., Dalla, C., Ferland-Beckham, C., Herremans, A., Karantzalos, K., Kas, M.J., Kokras, N., Parnham, M.J., Pavlidi, P., Pristouris, K., Steckler, T., Riedel, G., Emmerich, C.H., 2021. PEERS — An Open Science "Platform for the Exchange of Experimental Research Standards" in Biomedicine. Front. Behav. Neurosci. 15, 1–9. https://doi.org/10.3389/fnbeh.2021.755812

Sinclair, D., Featherstone, R., Naschek, M., Nam, J., Du, A., Wright, S., Pance, K., Melnychenko, O., Weger, R., Akuzawa, S., Matsumoto, M., Siegel, S.J., 2017. GABA-B Agonist Baclofen Normalizes Auditory-Evoked Neural Oscillations and Behavioral Deficits in the Fmr1 Knockout Mouse Model of Fragile X Syndrome. eNeuro 4. https://doi.org/10.1523/ENEURO.0380-16.2017

Smith, E.G., Pedapati, E. V., Liu, R., Schmitt, L.M., Dominick, K.C., Shaffer, R.C., Sweeney, J.A., Erickson, C.A., 2021. Sex differences in resting EEG power in Fragile X Syndrome. J. Psychiatr. Res. 138, 89–95. https://doi.org/10.1016/j.jpsychires.2021.03.057

Smith, N.J., Kutas, M., 2015. Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. Psychophysiology 52, 169–181. https://doi.org/10.1111/psyp.12320

Sohal, V.S., Rubenstein, J.L.R., 2019. Excitation-inhibition balance as a framework for investigating mechanisms in neuropsychiatric disorders. Mol. Psychiatry 24, 1248–1257. https://doi.org/10.1038/s41380-019-0426-0

Sohya, K., Kameyama, K., Yanagawa, Y., Obata, K., Tsumoto, T., 2007. GABAergic neurons are less selective to stimulus orientation than excitatory neurons in layer II/III of visual cortex, as revealed by in vivo functional Ca2+ imaging in transgenic mice. J. Neurosci. 27, 2145–2149. https://doi.org/10.1523/JNEUROSCI.4641-06.2007

Solomon, S.G., Kohn, A., 2014. Moving Sensory Adaptation beyond Suppressive Effects in Single Neurons. Curr. Biol. 24, R1012–R1022. https://doi.org/10.1016/j.cub.2014.09.001

Sørensen, E.M., Bertelsen, F., Weikop, P., Skovborg, M.M., Banke, T., Drasbek, K.R., Scheel-Krüger, J., 2015. Hyperactivity and lack of social discrimination in the adolescent Fmr1 knockout mouse. Behav. Pharmacol. 26, 733–740. https://doi.org/10.1097/FBP.0000000000000152

Spencer, C.M., Alekseyenko, O., Hamilton, S.M., Thomas, A.M., Serysheva, E., Yuva-Paylor, L.A., Paylor, R., 2011. Modifying behavioral phenotypes in Fmr1KO mice: genetic background differences reveal autistic-like responses. Autism Res. 4, 40–56. https://doi.org/10.1002/aur.168

Spencer, C.M., Graham, D.F., Yuva-Paylor, L.A., Nelson, D.L., Paylor, R., 2008. Social Behavior in Fmr1 Knockout Mice Carrying a Human FMR1 Transgene. Behav. Neurosci. 122, 710–715. https://doi.org/10.1037/0735-7044.122.3.710

Spencer, C.M., Serysheva, E., Yuva-Paylor, L.A., Oostra, B.A., Nelson, D.L., Paylor, R., 2006. Exaggerated behavioral phenotypes in Fmr1/Fxr2 double knockout mice reveal a functional genetic interaction between Fragile X-related proteins. Hum. Mol. Genet. 15, 1984–1994. https://doi.org/10.1093/hmg/ddl121

Stagg, C., Hindley, P., Tales, A., Butler, S., 2004. Visual mismatch negativity: the detection of stimulus change. Neuroreport 15, 487–491. https://doi.org/10.1097/01.wnr.00001

Stamberger, H., Nikanorova, M., Accorsi, P., Angriman, M., Benkel-herrenbrueck, I., Capovilla, G., Erasmus, C.E., Fannemel, M., Giordano, L., Helbig, K.L., Maier, O., Phalin, J., 2016. A neurodevelopmental disorder including epilepsy. Neurology 86, 954–962.

Stothart, G., Kazanina, N., 2013. Oscillatory characteristics of the visual mismatch negativity; what evoked potentials aren't telling us. Front. Hum. Neurosci. 7, 1–9. https://doi.org/10.3389/fnhum.2013.00426

Sulykos, I., Czigler, I., 2014. Visual mismatch negativity is sensitive to illusory brightness changes. Brain Res. 1561, 48–59. https://doi.org/10.1016/j.brainres.2014.03.008

Suzuki, T.H., Nunokawa, S., Jacobson, J.H., 1972. Visually evoked cortical response in light-adapted cat and liminal brightness discrimination. Jpn. J. Physiol. 22, 157–175.

Tackett, J.L., Brandes, C.M., King, K.M., Markon, K.E., 2019. Psychology's replication crisis and clinical psychological science. Annu. Rev. Clin. Psychol. 15, 579–604. https://doi.org/10.1146/annurev-clinpsy-050718-095710

Tada, M., Kirihara, K., Mizutani, S., Uka, T., Kunii, N., Koshiyama, D., Fujioka, M., Usui, K., Nagai, T., Araki, T., Kasai, K., 2019. Mismatch negativity (MMN) as a tool for translational investigations into early psychosis: A review. Int. J. Psychophysiol. 145, 5–14. https://doi.org/10.1016/j.ijpsycho.2019.02.009

Tauber, J.M., Vanlandingham, P.A., Zhang, B., 2011. Elevated levels of the vesicular monoamine transporter and a novel repetitive behavior in the Drosophila model of fragile X syndrome. PLoS One 6. https://doi.org/10.1371/journal.pone.0027100

Telias, M., 2019. Molecular Mechanisms of Synaptic Dysregulation in Fragile X Syndrome and Autism Spectrum Disorders. Front. Mol. Neurosci. 12, 51. https://doi.org/10.3389/fnmol.2019.00051

The Dutch-Belgian Fragile X Consortium, Bakker, C.E., Verheij, C., Willemsen, R., van der Helm, R., Oerlemans, F., Vermey, M., Bygrave, A., Hoogeveen, A.T., Oostra, B.A., Reyniers, E., De Boule, K., D'Hooge, R., Cras, P., van Velzen, D., Nagels, G., Martin, J.J., De Deyn, P.P., Darby, J.K., Willems, P.J., 1994. Fmr1 knockout mice: A model to study fragile X mental retardation. Cell 78, 23–33. https://doi.org/10.1016/0092-8674(94)90569-X

Thomas, A.M., Bui, N., Perkins, J.R., Yuva-Paylor, L.A., Paylor, R., 2012. Group 1 metabotropic glutamate receptor antagonists alter select behaviors in a mouse model for fragile X syndrome. Psychopharmacology (Berl). 219, 47–58. https://doi.org/10.1007/s00213-011-2375-4

Thurman, A.J., Potter, L.A., Kim, K., Tassone, F., Banasik, A., Potter, S.N., Bullard, L., Nguyen, V., McDuffie, A., Hagerman, R., Abbeduto, L., 2020. Controlled trial of lovastatin combined with an open-label treatment of a parent-implemented language intervention in youth with fragile X syndrome. J. Neurodev. Disord. 12, 1–17. https://doi.org/10.1186/s11689-020-09315-4

Thye, M.D., Bednarz, H.M., Herringshaw, A.J., Sartin, E.B., Kana, R.K., 2018. The impact of atypical sensory processing on social impairments in autism spectrum disorder. Dev. Cogn. Neurosci. 29, 151–167. https://doi.org/10.1016/j.dcn.2017.04.010

Till, S.M., Asiminas, A., Jackson, A.D., Katsanevaki, D., Barnes, S.A., Osterweil, E.K., Bear, M.F., Chattarji, S., Wood, E.R., Wyllie, D.J.A., Kind, P.C., 2015. Conserved hippocampal cellular pathophysiology but distinct behavioural deficits in a new rat model of FXS. Hum. Mol. Genet. 24, 5977–5984. https://doi.org/10.1093/hmg/ddv299

Tremblay, A., Newman, A.J., 2015. Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples. Psychophysiology 52, 124–139. https://doi.org/10.1111/psyp.12299

Turrigiano, G., 2011. Too many cooks? Intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. Annu. Rev. Neurosci. 34, 89–103. https://doi.org/10.1146/annurev-neuro-060909-153238

Tyzio, R., Nardou, R., Ferrari, D.C., Tsintsadze, T., Shahrokhi, A., Eftekhari, S., Khalilov, I., Tsintsadze, V., Brouchoud, C., Chazal, G., Lemonnier, E., Lozovaya, N., Burnashev, N., Ben-Ari, Y., 2014. Oxytocin-mediated GABA inhibition during delivery attenuated autism pathogenesis in rodent offspring. Science (80-.). 343, 675–680.

Vallone, F., Cintio, A., Mainardi, M., Caleo, M., Di Garbo, A., 2015. Existence of anticorrelations for local field potentials recorded from mice reared in standard condition and environmental enrichment. Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. 91, 1–15. https://doi.org/10.1103/PhysRevE.91.012702

Van der Aa, N., Kooy, R.F., 2020. GABAergic abnormalities in the fragile X syndrome. Eur. J. Paediatr. Neurol. 24, 100–104. https://doi.org/10.1016/j.ejpn.2019.12.022

Van der Molen, M.J.W., Van der Molen, M.W., Ridderinkhof, K.R., Hamel, B.C.J., Curfs, L.M.G., Ramakers, G.J.A., 2012a. Auditory change detection in fragile X syndrome males: A brain potential study. Clin. Neurophysiol. 123, 1309–1318. https://doi.org/10.1016/j.clinph.2011.11.039

Van der Molen, M.J.W., Van der Molen, M.W., Ridderinkhof, K.R., Hamel, B.C.J., Curfs, L.M.G., Ramakers, G.J.A., 2012b. Auditory and visual cortical activity during selective attention in fragile X syndrome: A cascade of processing deficiencies. Clin. Neurophysiol. 123, 720–729. https://doi.org/10.1016/j.clinph.2011.08.023

Van Diepen, H.C., Ramkisoensing, A., Peirson, S.N., Foster, R.G., Meijer, J.H., 2013. Irradiance encoding in the suprachiasmatic nuclei by rod and cone photoreceptors. FASEB J. 27, 4204–4212. https://doi.org/10.1096/fj.13-233098

Veeraragavan, S., Bui, N., Perkins, J.R., Yuva-Paylor, L.A., Carpenter, R.L., Paylor, R., 2011a. Modulation of behavioral phenotypes by a muscarinic M1 antagonist in a mouse model of fragile X syndrome. Psychopharmacology (Berl). 217, 143–151. https://doi.org/10.1007/s00213-011-2276-6

Veeraragavan, S., Bui, N., Perkins, J.R., Yuva-Paylor, L.A., Paylor, R., 2011b. The modulation of fragile X behaviors by the muscarinic M4 antagonist, tropicamide. Behav. Neurosci. 125, 783–790. https://doi.org/10.1037/a0025202

Veeraragavan, S., Graham, D., Bui, N., Yuva-Paylor, L.A., Wess, J., Paylor, R., 2012. Genetic reduction of muscarinic M4receptor modulates analgesic response and acoustic startle response in a mouse model of fragile X syndrome (FXS). Behav. Brain Res. 228, 1–8. https://doi.org/10.1016/j.bbr.2011.11.018

Verkerk, A.J.M.H., Pieretti, M., Sutcliffe, J.S., Fu, Y.H., Kuhl, D.P.A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F., Eussen, B.E., van Ommen, G.J.B., Blonden, L.A.J., Riggins, G.J., Chastain, J.L., Kunst, C.B., Galjaard, H., Thomas Caskey, C., Nelson, D.L., Oostra, B.A., Warran, S.T., 1991. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65, 905–914. https://doi.org/10.1016/0092-8674(91)90397-H

Vershkov, D., Benvenisty, N., 2017. Human pluripotent stem cells in modeling human disorders: The case of fragile X syndrome. Regen. Med. 12, 53–68. https://doi.org/10.2217/rme-2016-0100

Vinken, K., Vogels, R., Op de Beeck, H., 2017. Recent Visual Experience Shapes Visual Processing in Rats through Stimulus-Specific Adaptation and Response Enhancement. Curr. Biol. 27, 914–919. https://doi.org/10.1016/j.cub.2017.02.024

Voelkl, B., Altman, N.S., Forsman, A., Forstmeier, W., Gurevitch, J., Jaric, I., Karp, N.A., Kas, M.J., Schielzeth, H., Van de Casteele, T., Würbel, H., 2020. Reproducibility of animal research in light of biological variation. Nat. Rev. Neurosci. 21, 384–393. https://doi.org/10.1038/s41583-020-0313-3

Voelkl, B., Würbel, H., 2016. Reproducibility Crisis: Are We Ignoring Reaction Norms? Trends Pharmacol. Sci. 37, 509–510. https://doi.org/10.1016/j.tips.2016.05.003

Vogt, D., Cho, K.K.A., Lee, A.T., Sohal, V.S., Rubenstein, J.L.R., 2015. The Parvalbumin/Somatostatin Ratio Is Increased in Pten Mutant Mice and by Human PTEN ASD Alleles. Cell Rep. 11, 944–956. https://doi.org/10.1016/j.celrep.2015.04.019

von Kortzfleisch, V.T., Karp, N.A., Palme, R., Kaiser, S., Sachser, N., Richter, S.H., 2020. Improving reproducibility in animal research by splitting the study population into several 'mini-experiments.' Sci. Rep. 10, 1–16. https://doi.org/10.1038/s41598-020-73503-4

Voytek, B., Secundo, L., Bidet-caulet, A., Scabini, D., Shirley, I., Gean, A.D., Manley, G.T., Knight, R.T., 2010. Hemicraniectmy: A new model for human electrophysiology with high spatio-temporal resolution. J. Cogn. Neurosci. 22, 2491–2502. https://doi.org/10.1162/jocn.2009.21384.Hemicraniectomy

Wahlsten, D., Rustay, N.R., Metten, P., Crabbe, J.C., 2003. In search of a better mouse test. Trends Neurosci. 26, 132–136. https://doi.org/10.1016/S0166-2236(03)00033-X

Wahlstrom-Helgren, S., Klyachko, V.A., 2015. GABAB receptor-mediated feed-forward circuit dysfunction in the mouse model of fragile X syndrome. J. Physiol. 593, 5009–5024. https://doi.org/10.1113/JP271190

Walz, N., Mühlberger, A., Pauli, P., 2016. A Human Open Field Test Reveals Thigmotaxis Related to Agoraphobic Fear. Biol. Psychiatry 80, 390–397. https://doi.org/10.1016/j.biopsych.2015.12.016

Wan, F.J., Swerdlow, N.R., 1997. The basolateral amygdala regulates sensorimotor gating of acoustic startle in the rat. Neuroscience 76, 715–724. https://doi.org/10.1016/S0306-4522(96)00218-7

Warden, M.R., Cardin, J.A., Deisseroth, K., 2014. Optical neural interfaces. Annu. Rev. Biomed. Eng. 16, 103–129. https://doi.org/10.1146/annurev-bioeng-071813-104733

Watrous, A.J., Lee, D.J., Izadi, A., Gurkoff, G.G., Shahlaie, K., Ekstrom, A.D., 2013. A Comparative Study of Human and Rat Hippocampal Low-Frequency Oscillations During Spatial Navigation. Hippocampus 23, 656–661. https://doi.org/10.1002/hipo.22124

Wen, T.H., Lovelace, J.W., Ethell, I.M., Binder, D.K., Razak, K.A., 2019. Developmental Changes in EEG Phenotypes in a Mouse Model of Fragile X Syndrome. Neuroscience 398, 126–143. https://doi.org/10.1016/j.neuroscience.2018.11.047

Wenstedt, E.F.E., van Croonenburg, T.J., van den Born, B.J.H., Van den Bossche, J., Hooijmans, C.R., Vogt, L., 2021. The effect of macrophage-targeted interventions on blood pressure – a systematic review and meta-analysis of preclinical studies. Transl. Res. 230, 123–138. https://doi.org/10.1016/j.trsl.2020.11.002

Whitesell, J.D., Liska, A., Coletta, L., Hirokawa, K.E., Bohn, P., Williford, A., Groblewski, P.A., Graddis, N., Kuan, L., Knox, J.E., Ho, A., Wakeman, W., Nicovich, P.R., Nguyen, T.N., van Velthoven, C.T.J., Garren, E., Fong, O., Naeemi, M., Henry, A.M., Dee, N., Smith, K.A., Levi, B., Feng, D., Ng, L., Tasic, B., Zeng, H., Mihalas, S., Gozzi, A., Harris, J.A., 2021. Regional, Layer, and Cell-Type-Specific Connectivity of the Mouse Default Mode Network. Neuron 109, 545-559.e8. https://doi.org/10.1016/j.neuron.2020.11.011

Wilkinson, C.L., Nelson, C.A., 2021. Increased aperiodic gamma power in young boys with Fragile X Syndrome is associated with better language ability. Mol. Autism 12, 1–15. https://doi.org/10.1186/s13229-021-00425-x

Willemsen, R., Bontekoe, C.J.M., Severijnen, L.A., Oostra, B.A., 2002. Timing of the absence of FMR1 expression in full mutation chorionic villi. Hum. Genet. 110, 601–605. https://doi.org/10.1007/s00439-002-0723-5

Wilson, C., Rogers, J., Chen, F., Li, S., Adlard, P.A., Hannan, A.J., Renoir, T., 2021. Exercise ameliorates aberrant synaptic plasticity without enhancing adult-born cell survival in the hippocampus of serotonin transporter knockout mice. Brain Struct. Funct. 226, 1991–1999. https://doi.org/10.1007/s00429-021-02283-y

Witten, L., Oranje, B., Mørk, A., Steiniger-Brach, B., Glenthøj, B.Y., Bastlund, J.F., 2014. Auditory sensory processing deficits in sensory gating and mismatch negativity-like responses in the social isolation rat model of schizophrenia. Behav. Brain Res. 266, 85–93. https://doi.org/10.1016/j.bbr.2014.02.048

Wolfer, D.P., Stagljar-Bozicevic, M., Errington, M.L., Lipp, H.P., 1998. Spatial memory and learning in transgenic mice: Fact or artifact? News Physiol. Sci. 13, 118–123. https://doi.org/10.1152/physiologyonline.1998.13.3.118

Wurbel, H., 2000. Behaviour and the standardization fallacy. Nat. Genet. 26, 263. https://doi.org/10.1038/81541

Yan, T., Feng, Y., Liu, T., Wang, L., Mu, N., Dong, X., Liu, Z., Qin, T., Tang, X., Zhao, L., 2017. Theta oscillations related to orientation recognition in unattended condition: A vMMN study. Front. Behav. Neurosci. 11, 1–8. https://doi.org/10.3389/fnbeh.2017.00166

Yang, M., Silverman, J.L., Crawley, J.N., 2011. Automated three-chambered social approach task for mice. Curr. Protoc. Neurosci. https://doi.org/10.1002/0471142301.ns0826s56

Yau, S.Y., Chiu, C., Vetrici, M., Christie, B.R., 2016. Chronic minocycline treatment improves social recognition memory in adult male Fmr1 knockout mice. Behav. Brain Res. 312, 77–83. https://doi.org/10.1016/j.bbr.2016.06.015

Yizhar, O., Fenno, L.E., Prigge, M., Schneider, F., Davidson, T.J., Shea, D.J.O., Sohal, V.S., Goshen, I., Finkelstein, J., Paz, J.T., Stehfest, K., Fudim, R., Ramakrishnan, C., Huguenard, J.R., Hegemann, P., Deisseroth, K., 2011. Neocortical excitation / inhibition balance in information processing and social dysfunction. Nature 477, 171–178. https://doi.org/10.1038/nature10360

Yucel, G., McCarthy, G., Belger, A., 2007. fMRI reveals that involuntary visual deviance processing is resource limited. Neuroimage 34, 1245–1252. https://doi.org/10.1016/j.neuroimage.2006.08.050

Yuhas, J., Cordeiro, L., Tassone, F., Ballinger, E., Schneider, A., Long, J.M., Ornitz, E.M., Hessl, D., 2011. Brief report: Sensorimotor gating in idiopathic autism and autism associated with fragile X syndrome. J. Autism Dev. Disord. 41, 248–253. https://doi.org/10.1007/s10803-010-1040-9

Zeidler, S., Pop, A.S., Jaafar, I.A., de Boer, H., Buijsen, R.A.M., de Esch, C.E.F., Nieuwenhuizen-Bakker, I., Hukema, R.K., Willemsen, R., 2018. Paradoxical effect of baclofen on social behavior in the fragile X syndrome mouse model. Brain Behav. 8, e00991. https://doi.org/10.1002/brb3.991

Zhang, D., Yu, B., Liu, J., Jiang, W., Xie, T., Zhang, R., Tong, D., Qiu, Z., Yao, H., 2017. Altered visual cortical processing in a mouse model of MECP2 duplication syndrome. Sci. Rep. 7, 1–14. https://doi.org/10.1038/s41598-017-06916-3

Zhang, L.I., Bao, S., Merzenich, M.M., 2001. Persistent and specific influences of early acoustic environments on primary auditory cortex. Nat. Neurosci. 4, 1123–1130. https://doi.org/10.1038/nn745

Zhang, Y., Bonnan, A., Bony, G., Ferezou, I., Pietropaolo, S., Ginger, M., Sans, N., Rossier, J., Oostra, B., LeMasson, G., Frick, A., 2014. Dendritic channelopathies contribute to neocortical and sensory hyperexcitability in Fmr1-/y mice. Nat. Neurosci. 17, 1701–1709. https://doi.org/10.1038/nn.3864

Zhang, Y.Q., Bailey, A.M., Matthies, H.J.G., Renden, R.B., Smith, M.A., Speese, S.D., Rubin, G.M., Broadie, K., 2001. Drosophila fragile x-related gene regulates the MAP1B homolog Futsch to control synaptic structure and function. Cell 107, 591–603. https://doi.org/10.1016/S0092-8674(01)00589-X

Zhao, X., Gazy, I., Hayward, B., Pintado, E., Hwang, Y.H., Tassone, F., Usdin, K., 2019. Repeat instability in the fragile x-related disorders: Lessons from a mouse model. Brain Sci. 9. https://doi.org/10.3390/brainsci9030052

Zimmerman, M., Martinez, J.H., Young, D., Chelminski, I., Dalrymple, K., 2013. Severity classification on the Hamilton depression rating scale. J. Affect. Disord. 150, 384–388. https://doi.org/10.1016/j.jad.2013.04.028

Zupan, B., Sharma, A., Frazier, A., Klein, S., Toth, M., 2016. Programming social behavior by the maternal fragile X protein. Genes. Brain. Behav. 15, 578–587. https://doi.org/10.1111/gbb.12298

Zupan, B., Toth, M., 2008. Inactivation of the maternal fragile X gene results in sensitization of GABAB receptor function in the offspring. J. Pharmacol. Exp. Ther. 327, 820–826. https://doi.org/10.1124/jpet.108.143990

Zwetsloot, P.P., Van Der Naald, M., Sena, E.S., Howells, D.W., IntHout, J., De Groot, J.A.H., Chamuleau, S.A.J., MacLeod, M.R., Wever, K.E., 2017. Standardized mean differences cause funnel plot distortion in publication bias assessments. Elife 6, 1–20. https://doi.org/10.7554/eLife.24260

All supplementary files can be found in the online version of the paper:

https://www.sciencedirect.com/science/article/pii/S0149763422002111?via%3Dihub

Supplementary file 1 – Search string

Supplementary file 2 – Tests and outcomes table

Supplementary file 3 – Characteristics table

Supplementary file 4 – Risk of bias assessment

Supplementary file 5 – Forest plots

Supplementary file 6 – Publication bias funnel plots

Supplementary file 7 – Sensitivity analysis

Supplementary file 8 – Extracted data

Supplementary file 9 – Meta-analysis results

Supplementary file 10 – Subgroup statistics

Supplementary file 11 – Publication bias analysis statistics

Supplementary file 12 – PRI

5

Exploration of testing time and light as potential factors interacting with the behavioral phenotype of the Fmr1-KO mouse model.

María Arroyo-Araujo, Robbert Havekes, Peter Meerlo, Martien J.H. Kas.

Groningen Institute for Evolutionary Life Sciences, University of Groningen, The Netherlands

Manuscript in preparation for submission.

ABSTRACT

The value of animal models in biomedical research relies on their capacity to partly represent a human disorder and to predict an outcome, given specific baseline conditions. However, the translation of animal models into clinical trials has proven to be challenging given the high rate of conflicting results found in preclinical literature (Freedman et al., 2015; Gerlai, 2018; Voelkl & Würbel, 2021).

In a recent meta-analysis (MA), we showed large between-study variability regarding the Fmr1-Knockout (KO) behavioral phenotype. For several of the behavioral categories analyzed around 50% of the studies did not find a difference between the Fmr1-KO animals and the control group. Additionally, we ran subgroup analyses testing whether experimental factors could explain the heterogeneity of results (*e.g.,* animals' housing arrangement as single or group). Unfortunately, some of these analyses couldn't be performed as too few studies reported such details (less than 5 papers per behavioral category). Based on these findings, we decided to perform a behavioral experiment to investigate whether the phenotype differences between Fmr1 KO mice and their wild-type (WT) littermates could be influenced by environmental factors, specifically the time of testing relative to the light-dark cycle and the light conditions during the test. To this end, we subjected mice to four of the most frequently reported behavioral tests according to our meta-analysis, *i.e.*, the Elevated plus maze (EPM), Open field test (OF), Acoustic startle (ASR) response, and Pre-pulse inhibition (PPI). These tests measure some of the typical Fragile-X/autism-like symptoms. The tests were carried out early in the light or dark phase and animals were tested under bright or dim lighting conditions. With this study, we aimed to separate the influence of the lighting conditions *per se* from the possible circadian influence. Results suggest that the tests performed are differentially sensitive to the environmental conditions tested. For example, the ASR test detects an enhanced response during the light phase compared to the dark phase, while light condition in the test itself did not have any influence on the outcome. This result underscores the importance of more transparent reporting of methodology in pre-clinical studies, particularly given the impact on the interpretation and use of animal models, and their potential translation to clinical trials.

INTRODUCTION

The rigorous standardization of experimental procedures in animal research has been a practice advocated to reduce data variability within and between laboratories, increasing test sensitivity and reducing sample sizes needed (Richter et al., 2009). Despite this practice, scientific results -in specifically the field of behavioral phenotyping- have shown great variability between laboratories (Kafkafi et al., 2018; Paylor, 2009).

A preliminary insight on how rigorous standardization practices hamper preclinical data came from the debate on whether environmental enrichment, as a means to improve animals wellbeing, would reduce the precision and replicability of preclinical studies. Although initially some researchers were concerned that environmental enrichment would increase the variability of data at the cost of the reproducibility of results, it was later shown that housing enrichment does not affect the within-group variability and does not increase the risk of conflicting results across replicate studies (Wolfer et al., 2004). Moreover, biologically relevant housing enrichment has the benefit of animals being able to perform species-specific behaviors, which also helps to reduce the display of abnormal behaviors that may negatively affect test results (Van de Weerd et al., 2002).

Moreover, it is known that diverse environmental factors to which experimental animals are exposed during their development shape their phenotype expression later in life; this phenotypic flexibility is explained by the reaction norm (Schlichting, C.D. & Pigliucci, M., 1998). Briefly, the reaction norm describes how the interaction between genotypic and environmental factors (G x E) results in a range of possible phenotypes. In this sense, there is not a single possible phenotype for a population but a range of phenotypes that vary (with different likelihoods) according to the specific environmental and/or genetic characteristics of the population. Therefore, increasing the genotypic and/or environmental diversity of study population could be beneficial in terms of the replicability of results.

Thus, the commonly followed research practices are likely to generate (phenotypic) results that are specific and presumably narrow, and potentially difficult to replicate in different scenarios. Accordingly, if an experimental design were to diversify the experimental population, the results would be representative of the interaction between these diverse components (*i.e.,* GxE). For example, the inclusion of both female and male mice or different strain backgrounds and/or diverse environmental conditions or characteristics (*e.g.,* stimuli with different light intensities or diverse times of test) would make the results likely replicable and generalizable across the diverse conditions.

In this context, there have been recent reports on the effect of introducing systematic variation in preclinical studies. In one such study, Richter and colleagues made a simulation from existing data to compare low environmental variation against high environmental variation, according to environmentally standardized replicates or diversified study replicates (Richter et al., 2009). They found that the variance produced with standardized replicates was larger than with heterogenized replicates, indicating that rigorously standardizing environmental conditions do not minimize the variability between replicate studies. Moreover, they also reported that standardization led to 9.4% of false-positive results which was 8% more than with heterogenization. Similarly, Voelkl and colleagues, assessed the impact of heterogenized study design in the reproducibility of results (Voelkl et al., 2018). Based on meta-analyses results from preclinical studies, they simulated single-laboratory study design experiments and compared them to simulated multi-laboratory studies. They found that the simulated multi-lab study design resulted in more significant and accurate results compared to the single-lab studies in terms of how close each study design's estimated effect size was to the "true" overall effect size of the meta-analysis. They attributed this enhanced accuracy to the heterogeneity of the study samples captured by the multi-lab studies, indicating the importance of introducing variability in study samples to account for between-lab variability.

Although the benefits of introducing systematic variation seem promising, it is still unclear which environmental factors should be systematically varied to boost the replicability and external validity. Clearly, varying all possible factors would be logistically impossible and likely unnecessary. Nonetheless, it is suggested to first explore environmental factors that are biologically meaningful for the behavior to assess so they can affect the phenotypic expression (Voelkl et al., 2020).

In view of the potential effect of diverse environmental conditions to produce variable results, we decided to test whether the time of testing relative to light-dark cycle of the animals and the light conditions during the test itself could influence the phenotypic expression of the Fragile-X syndrome mouse model (Fmr1-KO). This decision was based on the observation of the large amount of conflicting results for locomotion, anxiety and acoustic sensory processing in the Fmr1-KO (Kazdoba et al., 2016; Melancia & Trezza, 2018) together with the results obtained from our metanalysis (MA; (Kat et al., 2022). Moreover, experimental details such as the time of testing and the light conditions during the test are rarely reported in studies so it is not possible to evaluate their influence on the behavioral outcome. However, both factors have previously been reported to affect the behavioral outcomes we aimed to assess. For instance, it has been reported that low color-temperature light at high intensity increases anxiety in the elevated plus maze (EPM) compared to lower intensity light. In contrast, the temperature and intensity

of light did not affect spontaneous locomotor activity in an open field (OF) arena (Kapogiannatou et al., 2016). As for the time of testing, a study found increased locomotor activity, decreased % prepulse inhibition, but no effect in anxiety in the light-dark box when animals were tested during the light phase compared to when they were tested in the dark phase (Richetto et al., 2019). These results suggest that the phase of the light-dark (LD) cycle and the actual light condition during testing could affect the behavioral outcomes for locomotion, anxiety, and acoustic processing. These observations underscore the importance of examining the potential influence of such environmental factors on behavioral outcomes in light of increasing within-study variability to potentially decrease between-lab data variability. Therefore, in the current study we aimed to investigate whether the LD-cycle phase and/or light condition during behavioral testing would result in variable phenotype expression across the different conditions in Fmr1 KO compared to WT littermates. Such diversity of results could partly explain the variability of result reported in the literature and our aforementioned MA. In addition, if the environmental factors under study would indeed influence the phenotypic expression, they could potentially be used to introduce systematic variation in these common behavioral tests. This could increase external validity of results, and hopefully increase their translational value of rodent models.

METHODS

All experiments were conducted under the Directive 2010/63/EU in accordance with the recommendations of the Guide for the Care and Use of Laboratory Animals, and approved by National Central Committee for scientific procedures on animals (CCD) and the Animal Welfare Body of the University of Groningen.

a. Animals

Fmr1-KO animals which do not produce FMRP protein nor have *FMR1* mRNA present (Mientjes et al., 2006) were bred *in-house* to generate male Fmr1-KO mice (referred to as KO) and wild-type male littermates (*i.e.*, WT), which served as control subjects. The line was maintained on a C57BL/6J background. Mice were weaned at 3-4 weeks of age, ear-clipped 2 weeks after birth to be genotyped by PCR and used for behavioral testing starting at 8-9 weeks of age.

According to a power calculation based on literature on the Fmr1-KO mice and the behavioral tasks that were planned, 40 KO and 40 WT experimentally naïve male mice were used for these experiments. Subjects were group housed in an enriched environment with 4 animals of mixed genotypes per cage. Light in the housing room was set

50 Lux and followed a 12:12 Light-Dark (LD) cycle. The housing room was set to ± 20 degrees Celsius and ~50% of relative humidity. Subjects had free access to water and food (Altronim), the animals were welfare-checked, weighed and cleaned weekly. The experimenter was blinded to the genotypes throughout the experiments.

Mice from the different cycle phases (light or dark) were housed in rooms opposite to each other with different L-D cycles so it was possible to test all animals during the light and dark phases of their cycles within the same day.

The experiment was conducted in three different batches separated by a month with the first batch starting in July, followed by August and September 2021. All batches had the same laboratory conditions and were all tested by the same experimenter although sample sizes were different as they depended on the number of offspring available.

Moreover, to corroborate the lack of Fmr1 expression in this mouse line, total cerebral RNA was isolated from a separate cohort of animals to be quantify by RT-PCR. Results confirmed that KO animals do not express Fmr1 (Appendix 1).

b. Behavioral testing

All tests were performed between one and two hours after the light transition (*i.e.,* early light phase or early dark phase; referred as 'cycle phase'). Animals were directly transferred from their home room to the test arena without acclimation to the test room and test environment, except for the ASR and PPI tests. The test order of the animals was block-randomized taking the location of the cage in the housing room as the blocking factor. All mice were tested only once in each behavioral test and always in the following order with one day to rest between tests: Elevated plus maze (EP), Open field (OF), Acoustic startle response (ASR) and Pre-pulse inhibition (PPI).

Behavioral assessment was performed under two different 'test lights': bright light (50 Lux) referred to as *Bright* or dim light (8 Lux) referred to as *Dim*. For the ASR and PPI tests, the *Dim* light intensity was total darkness to minimize modifications to the hardware setup. Subjects were tested in a counterbalanced manner according to the testing condition and this was varied across tests as much as possible.

i. Elevated Plus maze (EPM)

The EPM was used to assess anxiety-related behavior across genotypes; a longer relative time spent in open arms means the animal is less anxious. Two light gray mazes were used (Noldus Information Technology, Wageningen, the Netherlands) simultaneously (30 cm L x 6 cm W x 20 cm H, 50 cm above the floor). Sight from one maze to the other

was completely covered. Only mice coming for the same cage were tested at the same time. Subjects were placed in the center of the maze with the nose facing a random direction. Behavior of the mice was then recorded for 10 minutes. Time in open arms, closed arms and center was scored using EthoVision 16 (Noldus Information Technology, Wageningen, the Netherlands), to calculate the 'open arms ratio' [time in open arms/ (time in closed arms + time in open arms - time in center)], this is presented as a percentage in the results.

ii. Open field test (OF)

To assess spontaneous locomotion, two square (20 cm L x 20 W cm x 30 cm H) white arenas were used concurrently with cage mate mice. A mouse was placed in the center of the arena and video-recorded for 15 minutes. The outcome measure was the total distance traveled as tracked with EthoVision 16 (Noldus Information Technology, Wageningen, The Netherlands).

iii. Acoustic startle response (ASR)

Four SR-Lab startle response systems (ABS isolation cabinet 15" W x 14" D x 18" H; San Diego Instruments, Inc., San Diego, CA) were used to assess the startle response to auditory stimuli in both ASR and PPI experiments. The startle response was quantified by placing animals in a tubular enclosure coupled with an accelerometer sensor that reacts very quickly to sudden force changes. Both paradigms allowed for the animal to habituate to the box and background noise (65dB) for 5 minutes at the start of each paradigm. The ASR program consisted of presenting 12 different tones to the animals and record their startle response based on the measurement by the accelerometer. The tested tones were: no pulse (65 dB), 70, 75, 80, 85, 90, 95, 100, 105, 110, 115 and 120 dB. Each lasted for 40ms and was repeated 10 times in a pseudo-random way interspersed with inter-trial intervals (ITI) of 5-15sec. The outcome was the startle intensity (mV) to each tone presentations. Analysis was carried out with the raw data points (*i.e.,* data was not averaged). Due to a technical problem, data from one cohort of batch 2 was not usable and hence was excluded from the analysis (WT=2, KO=2).

iv. Pre-pulse inhibition (PPI)

The PPI paradigm consisted of presenting a pre-pulse paired with a startle pulse and record the startle response of the mouse. There were 3 different pre-pulse (PP) intensities, namely: 74, 80 & 86 dB, each of them lasted for 20 ms. After an inter-stimulus interval ISI of either 30 or 100 ms, a startle pulse (SP) of 120 dB was presented. The PPI paradigm consisted of 4 blocks. The first block was the 5 minutes of habituation period to the background noise (65dB). The second and fourth block presented the startle pulse (120 dB) to monitor habituation to the stimulus at the beginning and end of the experiment.

Block three consisted of the presentation of the different stimuli in single (e.g., only one startle or pre-pulse) and paired trails (i.e., PP + SP) with the different ISI durations, each of these combinations was presented 15 times in a pseudo-randomized order. Startle pulses lasted 40 ms while pre-pulses lasted 20 ms. The main outcome of this test was calculated with the following formula:

(1-(pre-pulse + startle pulse)/(startle pulse))*100

Where (pre-pulse + startle pulse) stands for the average response per subject to the paired trials, while the startle pulse represents the average of the single startle pulse trials excluding the trials presented in blocks 2 and 4. Due to the nature of this calculation, the analysis was carried with the average percentage of PPI (%PPI) per subject.

c. Data analysis

All data was extracted from their respective analysis software as excel files which were then imported into RStudio (v4.0.5 (R v3.6.3), Boston, MA, USA; lme4 package v1.1-27; ggplot2 package v3.3.5, multcomp package v1.4-16) to be analyzed with linear and mixed effects modeling to test the relationship of genotype, cycle phase, test light and batch with the predefined outcome measures. Using model comparison based on Akaike Information Criterion (AIC) the best model was selected. After selecting the optimal model specification, we applied model criticism by excluding all the observations with absolute residuals larger than 2 standard deviations above the mean (2% of the observations; see Baayen, 2008). The final models reflect the results based on the trimmed dataset. All models were validated via bootstrap sampling (R= 1000). The optimal models were chosen not only based on their goodness of fit to explain the data but also, they had to include the factor of interest (*i.e.,* genotype, cycle phase, test light) to answer the research question regardless of how informative they were in terms of the variance explained.

The optimal models according to our hypothesis found per test were the following:

EPM: Open arms ratio ~ Genotype + Cycle phase + Test light + (1|Batch)

OFT: Distance traveled ~ Genotype + Cycle phase + Test light + (1|Batch)

ASR: Startle ~ Genotype + Cycle phase + Test light + (1|Batch) + (1|Animal ID)
+ (1|Tone type)

PPI: %PPI ~ Genotype + Cycle Phase + Test light + (1|Batch) + (1|Animal ID)
+ (1| Tone type)

After visual inspection of residual plots, data that deviated from homoscedasticity or normality was converted with the natural logarithm (ASR, PPI) or squared-root (EPM, OF). Differences were considered to be statistically significant when $p \leq 0.05$.

Additional exploratory analysis was carried out to assess the effect of the batch as a fixed factor instead of random factor. The rest of the model remained the same unless otherwise specified. Chi-squared tests were used to determine the statistical significance of the effects.

Data is presented as mean ± standard deviation.

RESULTS

A. Elevated Plus Maze

This anxiety test showed a main effect on genotype where KO mice spent significantly ($\alpha = 0.05$) more time in the open arms compared to their WT counterparts ($|t| > 2$, $p < 0.05$; Appendix Table 1 & 2; Figure 1). This genotype effect was not influenced by the light/dark cycle phase nor by the test light ($|t| < 2$; Appendix Table 1 & 2). The goodness of fit of the model explained 16% of the variance in the data according to the R-squared component.

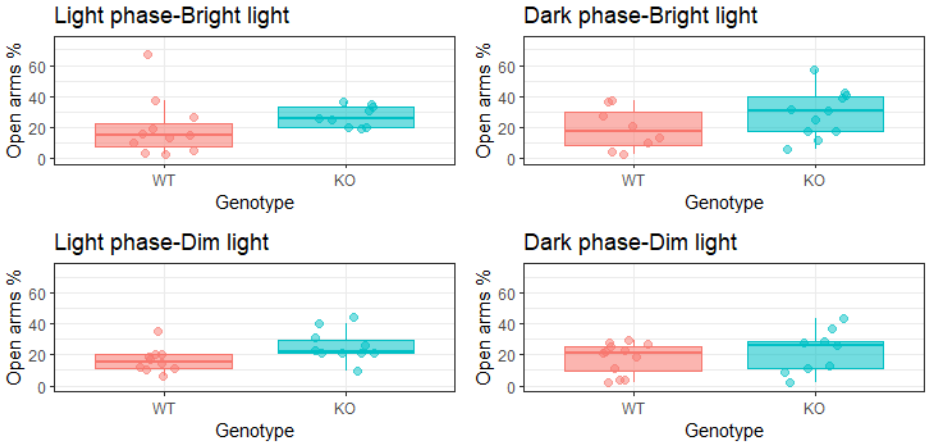


Figure 1. Percentage time spent in open arms in the Elevated plus maze across batches in the light phase (left panel) and dark phase (right panel), with testing lights bright (upper panel) or dim (lower panel). Red boxplots and symbols for WT mice and blue boxplots and symbols for KO mice.

In addition, the exploratory analysis indicated there were no differences between batches (|*t*|< 2; Appendix Table 3), nor an interaction between batch and other factors, which suggest a robust phenotype (Figure 2).

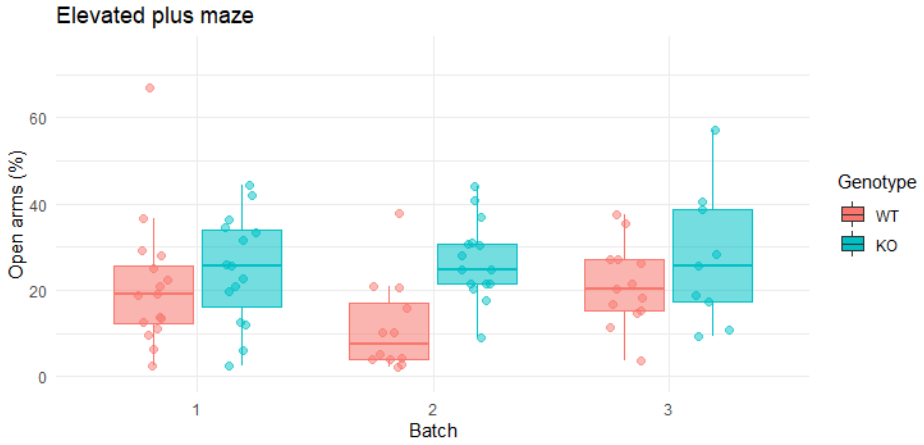


Figure 2. Percentage of open arms ratio in the Elevated plus maze pooled by batch. Red boxplots and symbols represent WT mice while blue boxplots and symbols are for KO mice.

a. Open Field

The distance traveled in the OF was used to assess spontaneous locomotor activity levels between genotypes. There was no difference found in the distance traveled between WT and KO mice (|*t*|< 2; Appendix Table 4 & 5). This finding was consistent regardless of whether the animals were tested in the light or dark phase and regardless of the test light in the arena (Figure 3). There were no interactions between genotype and the environmental factors. The accounted variance for this model was 11%.

The exploratory analysis on the batch effect showed a significant difference between the overall distance traveled between batch 1 and 3; batch 1 had a longer distance traveled (|*t*|> 2; Appendix table 6; Figure 4). The goodness-of-fit for this model was only 6% (Appendix Table 6).

In order to assess whether the locomotion outcome in the OFT could have been affected by previous experience in the EPM, an exploratory analysis assessed the total distance traveled in the 10 minutes of the EPM. This analysis revealed a significant interaction of the genotype with the test light where the KO animals moved more than WT animals (|*t*|>2) only when the test light was dim (Figure 5).

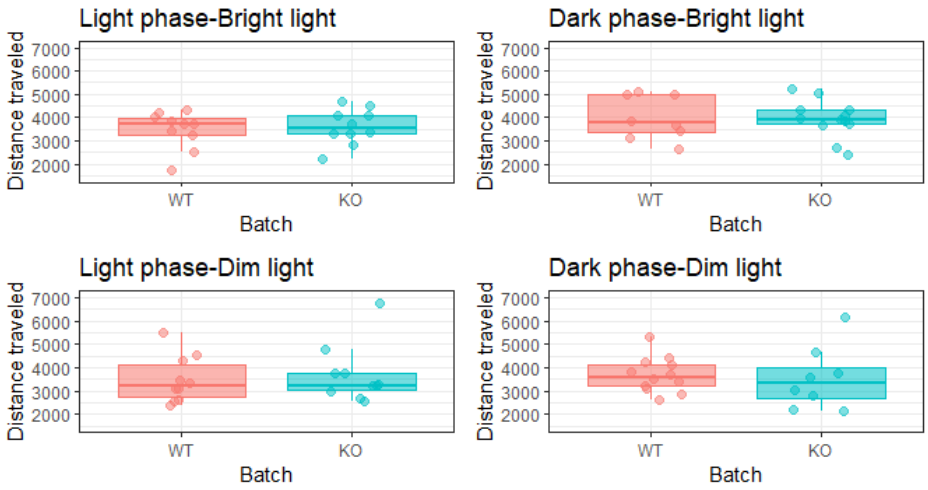


Figure 3. Distance traveled in the Open field across batches in the light phase (left panel) and dark phase (right panel), with bright lights (upper panel) or dim lights (lower panel). Red boxplots and symbols for WT mice and blue boxplots and symbols for KO mice.

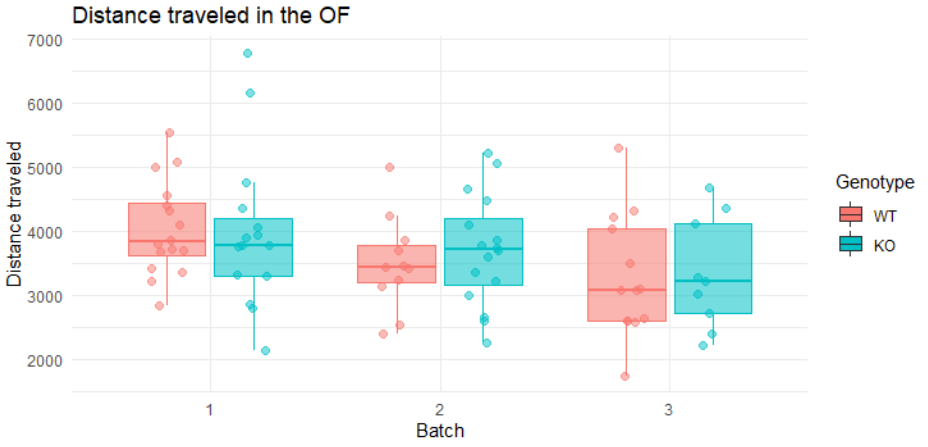


Figure 4. Distance traveled in the Open field pooled by batch. Red boxplots and symbols represent WT mice while blue boxplots and symbols are for KO mice.

b. Acoustic Startle Response

The ASR is commonly used to assess acoustic sensory processing. Here, there was no overall difference between genotypes ($|t| < 2$; Appendix table 7 & 8). As for the environmental factors assessed, a main effect on the phase of the LD-cycle was found; overall animals tested early in light phase startled more than animals tested early in the dark phase ($|t| > 2$; Appendix Table 7). The light condition during the test itself seemed to have no effect on the startle response. This model explained 52% of the variance. Moreover,

there were no interactions between the environmental conditions and the genotype (Figure 6).

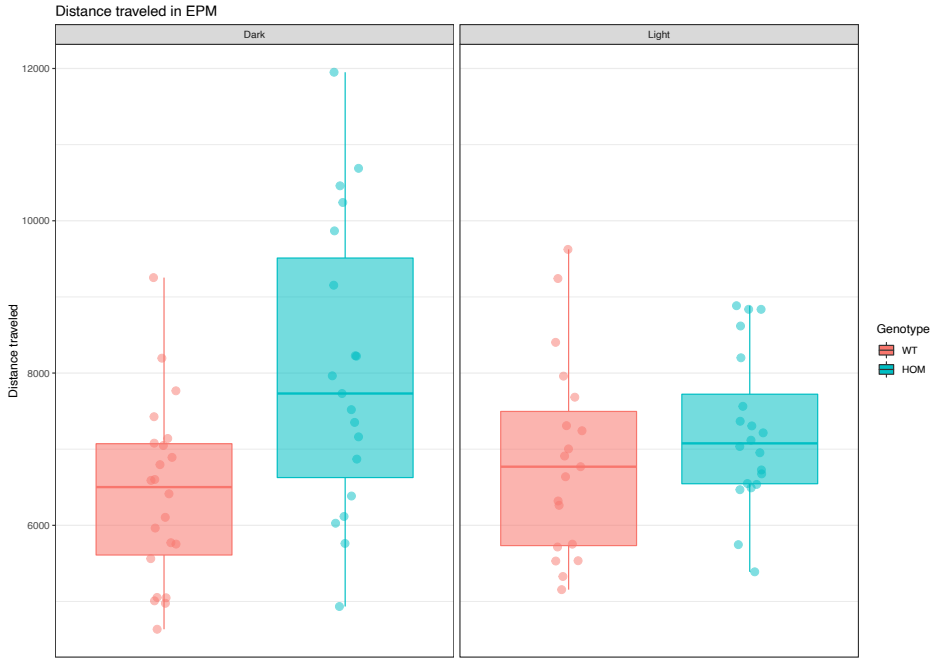


Figure 5. Total distance traveled in the EPM for WT mice (red) and Fmr1-KO (teal) for the Dim (left panel) and bright (right panel) test light conditions.

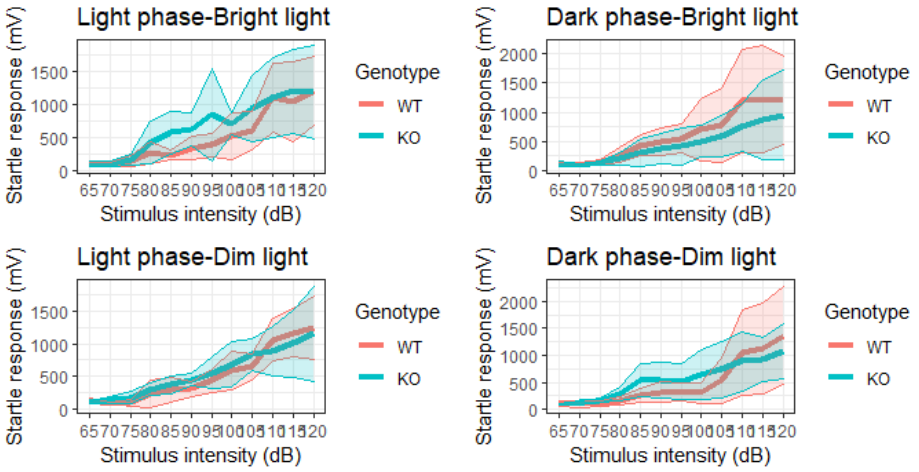


Figure 6. Acoustic startle response across different tone intensities by genotypes. Red line (average across individuals per genotype) and symbols (average across trials per animal) represent WT mice while blue line (average across individuals per genotype) and symbols (average across trials per animal) represent KO mice.

The exploratory analysis on the possible batch effect suggested that batch 2 was significantly different than the other two batches (|*t*|> 2; Appendix Table 9 & 10). However, when correcting for multiple comparisons of means with Tukey contrasts, this difference lost significance (Pr(>|z|) 0.08; Figure 7).

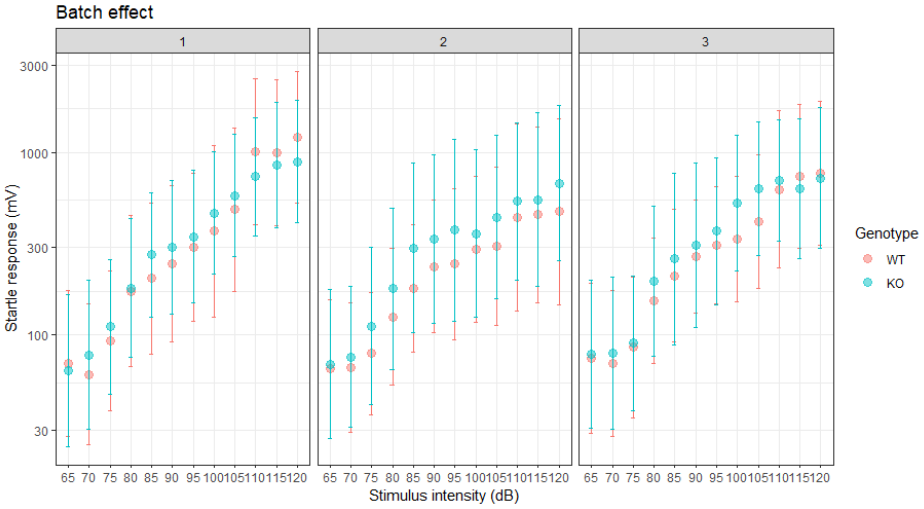


Figure 7. Startle response (mV) across the different stimuli intensities (65-120 dB) for WT (red symbols) and KO (teal symbols) animals belonging to the different experimental batches (panels 1 to 3).

c. Pre-Pulse Inhibition

The percentage of PPI was calculated to compare the sensory sensitivity between genotypes and conditions. There was no genotype effect nor a difference between the LD-cycle phase or the test light conditions for this test (|*t*|<2; Appendix Table 11 & 12). The variance explained by this model is 71% (Figure 8).

Two exploratory analyses were performed. The first one aimed to evaluate whether the different parameters used in the test would show pre-existing differences between genotypes that could have influence the percentage of PPI. Namely, the trials with single pre-pulse (PP), single startle pulse-only (SP) and the different durations of inter-stimulus intervals (ISI) during the paired trials were evaluated (Figure 9). For this analysis, the experimental environmental conditions were removed from the model since they showed to have no influence in the outcome measure.

The analysis revealed no pre-existing differences between genotypes for the 3 test parameters. For the pre-pulse only trials, a model trimming was performed of 36 residual outliers, the analyses performed after the model trimming showed no difference between genotypes (Figure 9, Panel A; |*t*|<2; Appendix Table 13 & 14). The trimmed data

frame was then used for the single startle pulse analysis which also showed no difference between genotypes (Figure 9, Panel B; $|t|<2$; Appendix Table 15 & 16). Similarly, the ISI showed no effect on the %PPI (Figure 9, Panel C; $|t|<2$; Appendix Table 17 & 18). In addition, the comparison of the different pre-pulses used revealed an overall effect on the stimuli intensity where all intensities where significantly different from each other after Tukey contrasts (Figure 9, Panel A; $Pr(>|z|) < 0.01$).

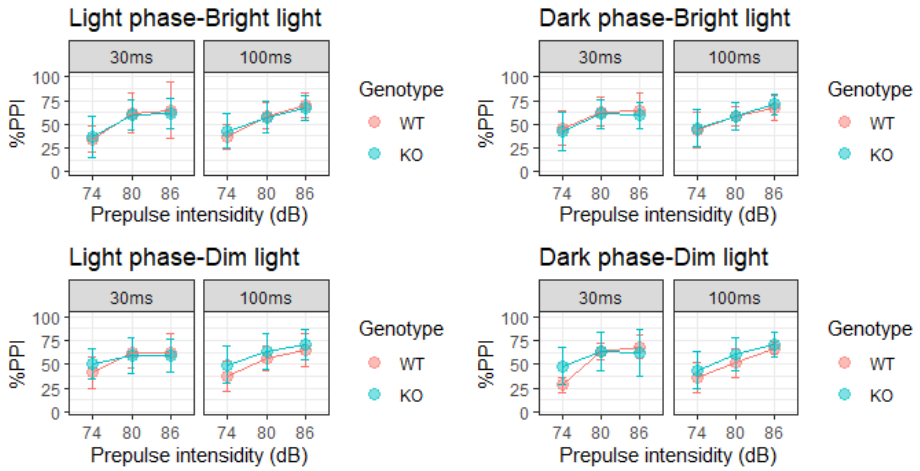


Figure 8. Percentage of PPI for the different genotypes across different pre-pulse intensities (74, 80 and 86 dB) with different ISI (30 ms or 100 ms). The different experimental conditions are expressed on top of each panel in the following order: Cycle phase -Test light condition. Red symbols represent WT animals, teal symbols represent KO animals.

The second exploratory analysis tested the difference between batches in which the experiment was performed (Figure 10). This analysis revealed no differences between batches (Figure 10; $|t|<2$; Appendix Table 19 & 20).

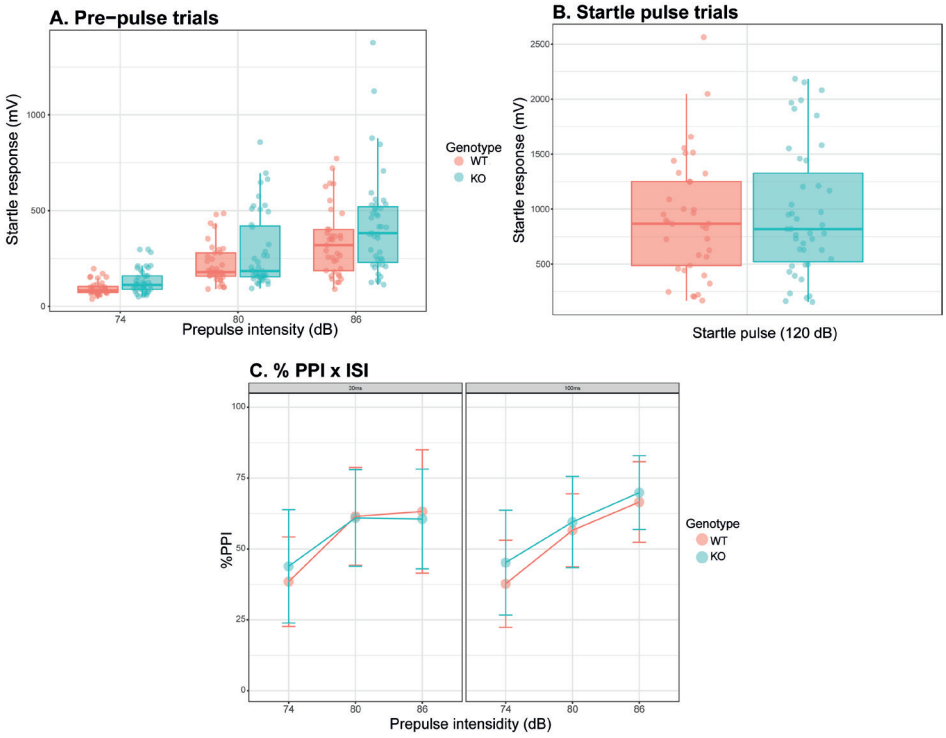


Figure 9. Exploratory results on the pre-pulse trials (A), the startle trials (B) and the different ISI durations (C) between genotypes. Red are WT animals; teal are KO animals. None of these analyses revealed difference between genotypes.

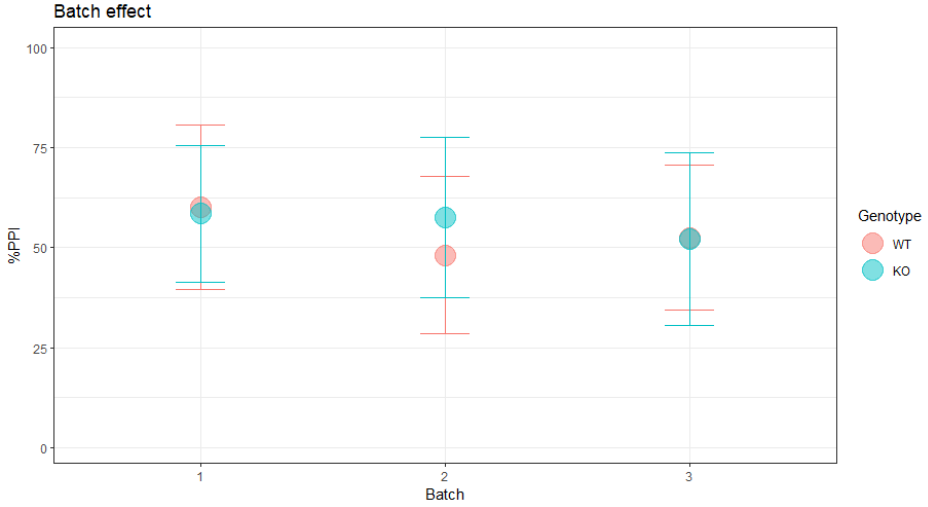


Figure 10. Percentage of pre-pulse inhibition across the different batches across genotypes. Red symbols for WT animals; teal symbols for KO animals.

DISCUSSION

The current study aimed to assess the influence of environmental factors on the behavioural phenotype expression in mice. In this way, we provide an initial attempt to identify factors that can increase the within study variability and the external validity of results. In the present study, the time of testing relative to the LD cycle and the test light in the arena were diversified to test their potential effect in behavioral outcomes for locomotion, anxiety, sensory processing and gating in the Fmr1-KO mouse model.

The main outcome measures for anxiety, locomotion and sensory gating showed no influence by neither of the environmental factors diversified in the study. Only the sensory processing outcome showed an effect for the time of testing. In addition, an exploratory analysis on the distance traveled in the EPM showed an interaction where the KO moved more than the WT only when the test light condition was bright. In contrast to the OFT findings, increased novelty-induced distance moved was observed in the EPM, the first paradigm in the sequence of experiments that were performed. This indicates that prior handling and testing may have influenced the novelty-induced hyperactivity that is usually reported for this KO model in the OFT (Kat et al., 2022). These diverse results may indicate that the environmental factors diversified in this experiment did not have enough influence on the outcome measures to change the phenotype expression. On the other hand, it underscores the influence of the testing order in the phenotype expression and, therefore, the importance of transparent reporting of behavioral studies.

Additional to the environmental influence, the genotype effect on behavior was also analyzed. No genotype effect was found for the OFT, ASR and PPI tests. These results are not representative of the genotype effects found in the corresponding meta-analyses; however, they do align with ~50% of the studies in the MA where no differences between the KO and WT animals were observed. This disparity of results may suggest that there are other environmental factors, besides time of testing and light in the test, that may have a stronger influence in the outcome measures here discussed; however, more transparent reporting is needed to be able to measure this in a meta-analysis.

CONCLUSION

The time of test relative to the LD-cycle affected the sensory processing in mice as was highlighted by the ASR test in the present study. On the other hand, the light intensity during the tests may have an influence in locomotion. Therefore, it is crucial that the reporting of environmental conditions of behavioral studies is improved so the po-

tential influence of such environmental condition in behavioral outcomes relevant for phenotype interpretation can be assessed. The batch effect seen in the OFT suggests that carrying out experiments in different batches could increase the within-experiment variability to potentially account for the between-experiments variability. More in-depth meta-analyses would be helpful to explore the possible influence of technical parameters in the phenotype expression. Moreover, the translational value of behavioral phenotypes that are only present in 50% of the studies in literature should be questioned if the cause of the diversity of results is not part of the phenotype, as it gives a false sense of model validity. Research practices that increase model validity (e.g., diversification of animal subjects and their environment) should be further studied and promoted to increase the value of preclinical studies.

REFERENCES

Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The Economics of Reproducibility in Preclinical Research. PLOS Biology, 13(6), e1002165. https://doi.org/10.1371/journal.pbio.1002165

Gerlai, R. (2018). Reproducibility and replicability in zebrafish behavioral neuroscience research. Pharmacology, Biochemistry, and Behavior. https://doi.org/10.1016/j.pbb.2018.02.005

Kafkafi, N., Agassi, J., Chesler, E. J., Crabbe, J. C., Crusio, W. E., Eilam, D., Gerlai, R., Golani, I., Gomez-Marin, A., Heller, R., Iraqi, F., Jaljuli, I., Karp, N. A., Morgan, H., Nicholson, G., Pfaff, D. W., Richter, S. H., Stark, P. B., Stiedl, O., … Benjamini, Y. (2018). Reproducibility and replicability of rodent phenotyping in preclinical studies. Neuroscience and Biobehavioral Reviews, 87, 218–232. https://doi.org/10.1016/j.neubiorev.2018.01.003

Kapogiannatou, A., Paronis, E., Paschidis, K., Polissidis, A., & Kostomitsopoulos, N. G. (2016). Effect of light colour temperature and intensity on τhε behaviour of male C57CL/6J mice. Applied Animal Behaviour Science, 184, 135–140. https://doi.org/10.1016/j.applanim.2016.08.005

Kat, R., Arroyo-Araujo, M., de Vries, R. B. M., Koopmans, M. A., de Boer, S. F., & Kas, M. J. H. (2022). Translational validity and methodological underreporting in animal research: A systematic review and meta-analysis of the Fragile X syndrome (Fmr1 KO) rodent model. Neuroscience & Biobehavioral Reviews, 139, 104722. https://doi.org/10.1016/j.neubiorev.2022.104722

Kazdoba, T. M., Leach, P. T., Yang, M., Silverman, J. L., Solomon, M., & Crawley, J. N. (2016). Translational Mouse Models of Autism: Advancing Toward Pharmacological Therapeutics. Current Topics in Behavioral Neurosciences, 28, 1–52. https://doi.org/10.1007/7854_2015_5003

Melancia, F., & Trezza, V. (2018). Modelling fragile X syndrome in the laboratory setting: A behavioral perspective. Behavioural Brain Research, 350, 149–163. https://doi.org/10.1016/j.bbr.2018.04.042

Mientjes, E. J., Nieuwenhuizen, I., Kirkpatrick, L., Zu, T., Hoogeveen-Westerveld, M., Severijnen, L., Rifé, M., Willemsen, R., Nelson, D. L., & Oostra, B. A. (2006). The generation of a conditional Fmr1 knock out mouse model to study Fmrp function in vivo. Neurobiology of Disease, 21(3), 549–555. https://doi.org/10.1016/j.nbd.2005.08.019

Paylor, R. (2009). Questioning standardization in science. Nature Methods, 6(4), 253–254. https://doi.org/10.1038/nmeth0409-253

Richetto, J., Polesel, M., & Weber-Stadlbauer, U. (2019). Effects of light and dark phase testing on the investigation of behavioural paradigms in mice: Relevance for behavioural neuroscience. Pharmacology Biochemistry and Behavior, 178, 19–29. https://doi.org/10.1016/j.pbb.2018.05.011

Richter, S. H., Garner, J. P., & Würbel, H. (2009). Environmental standardization: Cure or cause of poor reproducibility in animal experiments? Nature Methods, 6(4), 257–261. https://doi.org/10.1038/nmeth.1312

Schlichting, C.D. & Pigliucci, M. (1998). Phenotypic evolution: A reaction norm perspective. Sinauer Associates.

Van de Weerd, H. A., Aarsen, E. L., Mulder, A., Kruitwagen, C. L. J. J., Hendriksen, C. F. M., & Baumans, V. (2002). Effects of Environmental Enrichment for Mice: Variation in Experimental Results. Journal of Applied Animal Welfare Science, 5(2), 87–109. https://doi.org/10.1207/S15327604JAWS0502_01

Voelkl, B., Altman, N. S., Forsman, A., Forstmeier, W., Gurevitch, J., Jaric, I., Karp, N. A., Kas, M. J., Schielzeth, H., Van de Casteele, T., & Würbel, H. (2020). Reproducibility of animal research in light of biological variation. Nature Reviews Neuroscience, 21(7), 384–393. https://doi.org/10.1038/s41583-020-0313-3

Voelkl, B., Vogt, L., Sena, E. S., & Würbel, H. (2018). Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLOS Biology, 16(2), e2003693. https://doi.org/10.1371/journal.pbio.2003693

Voelkl, B., & Würbel, H. (2021). A reaction norm perspective on reproducibility. Theory in Biosciences, 140(2), 169–176. https://doi.org/10.1007/s12064-021-00340-y

Wolfer, D. P., Litvin, O., Morf, S., Nitsch, R. M., Lipp, H.-P., & Würbel, H. (2004). Laboratory animal welfare: Cage enrichment and mouse behaviour. Nature, 432(7019), 821–822. https://doi.org/10.1038/432821a

APPENDIX

qPCR for Fmr1-KO confirmation

Total cerebral RNA was isolated from 4 HET-WT and 3 HET-KO mice with ages varying from 5-21 weeks. The brains were homogenized using Trizol in the TissueLyser II (Qiagen). 1 µg of total RNA from each sample was used in a RT reaction with the RevertAid RT Kit (#K1691, Thermo Scientific), oligo(dT) 18 mix and dNTP mix (10mM). Quantitative RT-PCR reactions were performed in the ABI 7300 Real-Time PCR System with SYBR Green PCR Master Mix (Applied Biosystems). The same primers as previously reported were used (FMR1: forward primer: 5V-CCGAACAGATAATCGTCCACG, reverse primer: 5V-ACGCTGTCTGGCTTTTCCTTC; GAPDH: forward primer: 5V- CCTGGAGAAACCTGCCAAGTAT, reverse primer: 5V-CCCTCAGATGCCTGCTTCA) (Mientjes et al., 2006). The level of Fmr1 expression were calculated relatively to the GAPDH expression and normalized to the average expression in WT animals. The CT value of GAPDH was approximately 17 in all samples. Results from this analysis confirmed the lack of Fmr1 expression in the KO mice.

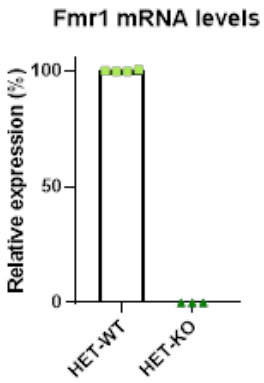


Table 1. EPM Fixed effects

	Estimate	Std. Error	t value	Pr(>Chisq)
(Intercept)	3.846544	0.306275	12.559	< 2.2e-16
Genotype: KO	1.058679	0.296400	**3.572**	0.0003546
Light phase: Light	-0.001463	0.295436	-0.005	0.9960491
Test light: Light	0.139581	0.296107	0.471	0.6373644

R-squared: **0.1632789**

Table 2. EPM Random Effects

Groups	Name	Variance	Std. Deviation
Batch	(Intercept)	0.03478	0.1865
Residual		1.72060	1.3117

Number of observations: 79, groups: Batch, 3

Table 3. EPM exploratory analysis on batches

	Estimate	Std. Error	t value	Pr(>|t|)	BootMed
(Intercept)	3.89353	0.35711	10.903	< 2e-16	3.89252
Genotype: KO	0.99199	0.31226	**3.177**	0.00217	1.00359
Light phase: Light	0.26596	0.31066	0.856	0.39470	0.27362
Test light: Light	0.08452	0.31042	0.272	0.78617	0.10815
Batch 2	-0.41593	0.36547	-1.138	0.25876	-0.42317
Batch 3	0.24125	0.38741	0.623	0.53539	0.23281

Adjusted R-squared: **0.09030**

Table 4. OFT fixed effects

	Estimate	Std. Error	t value	Pr(>Chisq)
(Intercept)	60.0862	1.9466	30.867	<2e-16
Genotype: KO	-0.3697	1.5964	-0.232	0.8169
Light phase: Light	-2.3135	1.5774	-1.467	0.1425
Test light: Light	1.8877	1.5875	1.189	0.2344

R-squared: **0.1178939**

Table 5. OFT random effects

Groups	Name	Variance	Std. Deviation
Batch	(Intercept)	4.306	2.075
Residual		48.970	6.998

Number of observations: 79, groups: Batch, 3

Table 6. OFT exploratory analysis on batches

	Estimate	Std. Error	t value	Pr(>|t|)	BootMed
(Intercept)	62.5033	1.7974	34.774	<2e-16	62.47874
Genotype: KO	-0.4195	1.6014	-0.262	0.7941	-0.35373
Light phase: Light	-2.2352	1.5784	-1.416	0.1610	-2.22936
Test light: Light	1.8481	1.5885	1.163	0.2485	1.95217
Batch 2	-2.4660	1.8684	-1.320	0.1910	-2.48640
Batch 3	-5.0058	1.9712	**-2.539**	0.0132	-5.07747

R-squared: **0.06543842**

Table 7. ASR fixed effects

	Estimate	Std. Error	t value	Pr(>Chisq)
(Intercept)	5.38166	0.26876	20.024	< 2e-16
Genotype: KO	0.14208	0.08018	1.772	0.07638
Light phase: Light	0.16404	0.08007	**2.049**	0.04048
Test light: Light	0.08213	0.08009	1.026	0.30510

R-squared: **0.5320592**

Table 8. ASR random effects

Groups	Name	Variance	Std. Deviation
Animal ID	(Intercept)	0.114060	0.33773
Tone type	(Intercept)	0.765335	0.87483
Batch	(Intercept)	0.006382	0.07989
Residual		0.740312	0.86041

Number of observations: 8981, groups: ID, 76; Type, 12; Batch, 3

Table 9. ASR fixed effects for exploratory analysis on batches

	Estimate	Std. Error	t value
(Intercept)	5.46610	0.26711	20.464
Genotype: KO	0.15064	0.08016	1.879
Light phase: Light	0.15929	0.07986	1.995
Test light: Light	0.09065	0.07993	1.134
Batch 2	-0.20634	0.09630	**-2.143**
Batch 3	-0.07477	0.09645	-0.775

Table 10. ASR random effects for exploratory analysis on batches

Groups	Name	Variance	Std. Deviation
Animal ID	(Intercept)	0.1125	0.3354
Tone type	(Intercept)	0.7569	0.8700
Residual		0.7526	0.8675

Number of observations: 9076, groups: ID, 76; Type, 12

Table 11. PPI fixed effects

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	55.7703	6.1481	11.9704	9.071	1.03e-06
Genotype: KO	-0.1578	3.4955	73.3177	-0.045	0.964
Cycle phase: Light	-2.2659	3.4681	72.8651	-0.653	0.516
Test light: Light	-0.7760	3.4562	72.6494	-0.225	0.823

R-squared: **0.7212363**

Table 12. PPI random effects

Groups	Name	Variance	Std. Deviation
Animal ID	(Intercept)	204.307	14.294
Tone type	(Intercept)	137.212	11.714
Batch	(Intercept)	4.801	2.191
Residual		154.982	12.449

Number of observations: 460, groups: ID, 78; Trial, 6; Batch, 3

Table 13. Fixed effects on the exploratory analysis of pre-pulse-only trials across genotypes

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.82043	0.07390	85.70914	65.228	< 2e-16
Genotype: KO	0.19085	0.09801	77.91114	1.947	0.05510
Pre-pulse 80dB	0.79814	0.03860	41.97556	20.678	< 2e-16
Pre-pulse 86dB	-0.11096	0.03862	42.05690	-2.873	0.00634

Table 14. Random effects on the exploratory analysis of pre-pulse intensities

Groups	Name	Variance	Std. Deviation
Animal ID	(Intercept)	0.17490	0.4182
Tone type	(Intercept)	0.01349	0.1161
Residual		0.73798	0.8591

Number of observations: 3744, groups: ID, 80; Tone type, 45

Table 15. Fixed effect of the exploratory analysis of the startle pulse across genotypes

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.21198	0.07940	102.66604	53.048	<2e-16
Genotype: KO	0.19188	0.09828	77.91542	1.952	0.0545
Pre-pulse 80dB	0.69494	0.05126	41.85578	13.557	<2e-16
Pre-pulse 86dB	1.13058	0.05127	41.87239	22.053	<2e-16

Table 16. Random effects of the exploratory analysis of the startle pulse

Groups	Name	Variance	Std. Deviation
Animal ID	(Intercept)	0.17603	0.4196
Trial	(Intercept)	0.01088	0.1043
Residual		0.73493	0.8573

Number of observations: 3741, groups: ID, 80; Trial, 45

Table 17. Fixed effects on the exploratory analysis on ISI durations across genotypes

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	52.9843	8.0955	5.1118	6.545	0.00115
Genotype: KO	0.4666	3.4361	77.8230	0.136	0.89234
ISI 100ms	1.3817	10.7296	4.0000	0.129	0.90375

Table 18. Random effects on the exploratory analysis on ISI durations across genotypes

Groups	Name	Variance	Std. Deviation
Animal ID	(Intercept)	205.500	14.335
Tone type	(Intercept)	170.691	13.065
Batch	(Intercept)	6.034	2.456
Residual		159.668	12.636

Number of observations: 480, groups: ID, 80; Tone type, 6; Batch, 3

Table 19. Fixed effects of the exploratory analysis on batches

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	59.1440	6.3635	14.3711	9.294	1.85e-07
Genotype: KO	-0.1227	3.5332	72.0000	-0.035	0.972
Test light: Bright	-0.4416	3.4812	72.0000	-0.127	0.899
Cycle phase: Light	1.9318	3.4960	72.0000	-0.553	0.582
Batch 2	-5.5864	4.0366	72.0000	-1.384	0.171
Batch 3	-6.3647	4.4098	72.0000	-1.443	0.153

Table 20. Random effects of the exploratory analysis on batches

Groups	Name	Variance	Std. Deviation
Animal ID	(Intercept)	206.6	14.37
Trial	(Intercept)	138.6	11.77
Residual		159.4	12.63

Number of observations: 468, groups: ID, 78; Trial, 6

The perks of a Quality System in academia

María Arroyo-Araujo and Martien J.H. Kas

Groningen Institute for Evolutionary Life Sciences, University of Groningen, the Netherlands

Published in Neuroscience Applied (Volume 1, 2022, 100001).

This commentary is linked to the article in Appendix 1 of this thesis.

In recent years, the biomedical research community has reappraised the importance of rigorous research practices for delivering reliable data that can support the development of novel, safe and effective therapies [1, 2].

With the help of meta-analyses, it has become evident that the way preclinical studies are planned, conducted, and reported are often far from optimal and in some cases might even hamper the drug development pipeline. Published studies with poor validity (internal, external, and of construct) and vague reporting mislead researchers to build upon this flawed knowledge; this, in turn, slows down the development of new drugs as it inflates treatment efficacy expectations [3] and contributes to failed clinical trials. Furthermore, it has been suggested that lack of rigorous research practices and report transparency, together with our shortage of understanding of biological variables that explain between laboratory variation contribute to the low turnover of preclinical studies [4, 8] which generates ethical concerns regarding the use of animals and public resources.

In an attempt to develop strategies for improving research practices in biomedical research, the European Quality In Preclinical Data (EQIPD) consortium was created through the European Union's Innovative Medicine Initiative (IMI). As part of the overall project, the EQIPD consortium has developed and released for public use a quality system (QS) meant to support the generation of robust and reliable data by promoting rigorous research practices [4].

The EQIPD-QS' approach focuses on the design, conduct, analysis, and report of unbiased studies to deliver robust results and thus, contribute to the development of effective and safe therapeutic targets. Moreover, the system was designed to be applied in academic and industry research contexts.

Formal quality management is a solution commonly applied in industry. In academia, there are rather few examples of a quality management system applied to support basic and applied research. As we will illustrate below, the benefits of applying a user-friendly quality system such as the EQIPD-QS may be as obvious even in an academic environment as relevant.

This is especially true considering the somewhat underestimated large academic input into the drug discovery pipeline. This has become more evident over the years as indicated by the increase of FDA-approved drugs released by academia: from 24% in 1998-2007, up to 48% in 2011, and 55% in 2015 [5, 6]. Moreover, this increment has been

achieved while pharma companies keep disappearing or decreasing their manpower, emphasizing the need for a quality system suitable for an academic context.

Thereby, the implementation of a QS such as EQIPD's could promote the formation of more conscientious researchers who can establish clear and efficient communication between labs, in and out of academia, while creating an integral scientific setting and improving the research culture in preclinical research. Therefore, our research group at the University of Groningen (the Netherlands) was one of the first to start implementing the EQIPD Quality System; this commentary aims to highlight important lessons learned.

A key factor for the successful implementation and maintenance of a quality system is a clear understanding of its purpose. Thus, we have formulated the reasons that motivated the implementation of the EQIPD- QS in our lab to which the reader might identify with.

Table 1. Reasons to implement EQIPD-QS in a lab of the University of Groningen.

The burden of implementing a QS was minimized since The Netherlands has rather strict regulations regarding the use of animals for research so we felt that the QS would not imply much effort on top of our standard administration.
Academic research in Europe has become highly collaborative as diverse consortia illustrate. There is a need to facilitate the communication between partners to ensure the vision and goals of the project are achieved on time in full; collaborators must be able to rely on each other's performance, including scientific integrity matters.
One of the main tasks of a university is to train students and new researchers and to provide the best environment for this. If a QS promotes rigorous research practices to produce robust and meaningful data then, implementing such a QS would help our group to train pre and postgraduate students and better prepare them for a career as independent scientists.
While we used to think of research integrity as being related to falsification, fabrication, and plagiarism, current definitions of scientific misconduct adopted by The European Federation of Academies of Science and Humanities (ALLEA) and various universities have expanded and now include "good data practices" such as those related to data management and storage, improper research design, among others [7].
There is a constantly increasing bureaucratic burden that is difficult to manage especially for the head of the lab who, in addition to management and supervision of students and staff, is expected to travel to meetings and collaborative partners, sits on various committees, etc. Having a systematic approach to oversee the lab's activities and progress is what also contributed to our decision.

The following part of this communication provides a brief overview of our experience implementing the EQIPD- QS, followed by the lessons learned.

In the case of our institute, as good research practices were already in place, the QS mainly served as a guide to assemble puzzle pieces. The documents that were needed

to implement the QS *core requirements,* listed in Table 2, were divided into two main categories related to university general regulations or local lab guidelines.

Table 2. The 18 QS *core requirements* and approximate time spent in each of them

Core Requirement	**Time taken to complete**				**Described in a QS stand-alone document**
	No time (the item was covered by an already existing document)	*Minutes (around 30 min)*	*Hours (4-5 hours)*	*Days (3-4 working days)*	
Process owner must be identified for the Quality System		X			
Communication process must be in place	X	X			
The research unit must have defined quality objectives			X		
All activities must comply with relevant legislation and policies	X				
The research unit must have a procedure to act upon concerns of potential misconduct		X			
Generation, handling, and changes to data records must be documented				X	
Data storage must be secured at least for as long as required by legal, contractual, or other obligations or business needs	X				
Reported research outcomes must be traceable to experimental data	X				
Reported data must disclose all repetitions of the test regardless of the outcome	X				
Investigator must declare in advance whether a study is intended to inform a formal knowledge claim		X			
All personnel involved in research must have adequate training and competence to perform assigned tasks		X			
Protocols for experimental methods must be available	X				
Adequate handling and storage of samples and materials must be ensured	X				

Table 2. The 18 QS *core requirements* and approximate time spent in each of them (continued)

Core Requirement	Time taken to complete				Described in a QS stand-alone document
	No time (the item was covered by an already existing document)	*Minutes (around 30 min)*	*Hours (4-5 hours)*	*Days (3-4 working days)*	
Research equipment and tools must be suitable for the intended use and ensure data integrity	X				
Risk assessment must be performed to identify factors affecting the generation, processing, and reporting of research data		X			
Critical incidents and errors during study conduct must be analyzed and appropriately managed		X			
An approach must be in place to monitor the performance of the EQIPD Quality System, and address identified issues		X			
Resources for sustaining the EQIPD Quality System must be available		X			

The university regulations comprise records related to general topics relevant for preclinical studies such as the research code of conduct, the access and use of a data archiving system, the institute licenses for the use of animals and certain drug compounds, among others. In general, the content of these regulations hardly ever changes, and therefore these documents only had to be compiled once the system was adopted. In addition, since animal research regulations are so strict in The Netherlands this step only implied entering the already existing records to the QS *dossier* directory, so no further action was required.

As for the lab regulations, these documents and records address specific aspects of how research is planned and performed and kept up-to-date albeit with the changing staff. This includes the training of personnel, the compilation of standard operating procedures (SOP's), ethical approval protocols to perform specific experiments, etc. In some cases, this content had to be adapted to the templates provided by the QS, which made it clear and accessible, and also made any necessary later updates almost effortless.

Once all documents were identified, updated and/or created, they were sorted out following the system's guidance and recommendations until the completion of the

implementation. The time taken to complete the *core requirements* of the QS is listed in table 2.

When the implementation was finalized, a formal assessment meeting was carried out to evaluate the fitness of the QS implementation. Some documents were created during the implementation and together with lab reports they were shared in advance with the assessment team, while an overview of the expectations of the implementation was sent to us. There were two 2-hour meetings in which a team of 3 assessors and two staff members from the university lab went through a checklist together. This checklist was focused on reviewing the fulfilment and fitness of the QS *core requirements* to our specific laboratory setting. In some instances, discussion and examples were exchanged between both parties to make sure the QS core requirements would hold valid for the different types of experiments and scenarios in our lab. At the end of the formal assessment, a series of recommendations to improve the fitness of the implementation were sent to us; these aimed to further promote and follow-up a positive research culture in our lab via day-to-day research practices overseen by the QS.

During the QS implementation, we came across different scenarios and we think the lessons learned from these are worth sharing.

Lesson # 1 - The devil is not so bad as he is painted

The time invested in the implementation of the quality system was limited since the lab was already operating at a high level. The implementation of EQIPD-QS does not necessarily result in more work where high standards are met while it strongly supports the generation of robust data.

Lesson # 2 – Quality System is a self-reflection tool

By following the step-by-step approach of the QS we identified unintended gaps in our training. In the case of our lab, MSc and Ph.D. students have a mandatory course addressing research integrity topics; however, there is no such course for Postdocs. This was easily solved by making the relevant university guidelines more easily accessible for all lab members to be properly informed.

Lesson # 3 – Do not hesitate to ask for help

As a university lab, we are required to archive all research data in the university repository. However, the administration of the repository entrusts the lab manager with approval to

edit records in case of incomplete/mistaken back-ups; this goes against EQIPD-QS recommendations. Given that the university archive regulations go beyond any lab's reach, we contacted the department of Information Communication and Technology (ICT) and they easily modified the read/write permissions for our lab members. Solutions turned out to be much easier than feared and anticipated, and we only needed to ask for help.

Lesson # 4 – Facilitation of the onboarding process

Onboarding new employments in a research team and institute usually requires time. Having the QS in place provides a step-by-step guided process with documentation that can be followed by the new employee without overlooking important local, national, and/or international procedures and regulations. Moreover, interaction among co-workers within the research team who follow the same steps facilitates the onboarding process further by having a feedback system already in place.

Personally, the implementation of EQIPD-QS made me aware of the urgent need to change the research culture for preclinical studies.

The familiarity with concepts like randomization and blinding shadows their importance on the eyes of scientists that easily forget to put them in practice and/or report whether they were carried out. Without specific guidelines on part of journals, missing items like these may go unnoticed until after publication; by then, it is difficult to assess the validity of the study and the worth of the resources invested. Likewise, published data with a high risk of bias tend to have a lower weight in meta-analysis studies, further limiting the contribution of the study.

In summary, implementing a QS such as EQIPD's in academia can promote the development of habits that boost the quality of executing and reporting preclinical research. We hope this empirical report will encourage fellow researchers to change the generally accepted way studies are usually conducted and reported so more meaningful results can be achieved in preclinical sciences.

ACKNOWLEDGMENTS

We would like to thank Peter Meerlo and Robbert Havekes for the valuable input provided to this commentary, as well as to Anton Bespalov, Björn Gerlach, and the assessment team for the very much appreciated guidance throughout the QS implementation process.

FUNDING

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777364. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

DECLARATION OF INTEREST

None

REFERENCES

[1] Bishop, D. (2019). Rein in the four horsemen of irreproducibility. Nature, 568(7753), 435–435.

[2] Giles, J. (2006). Animal experiments under fire for poor design. Nature, 444(7122), 981–981.

[3] Sena, E. S., Currie, G. L., McCann, S. K., Macleod, M. R., & Howells, D. W. (2014). Systematic Reviews and Meta-Analysis of Preclinical Studies: Why Perform Them and How to Appraise Them Critically. Journal of Cerebral Blood Flow & Metabolism, 34(5), 737–742.

[4] Bespalov, A., Bernard, R., Gilis, A., Gerlach, B., Guillen, J., Castagne, V., Lefevre, I., Ducrey, F., Monk, L., Bongiovanni, S., Altevogt, B., Arroyo Araujo, M., Bikovski, L., de Bruin, N., Castaños-Vélez, E., Dityatev, A., Emmerich, C. H., Fares, R., Ferland-Beckham, C., … Steckler, T. (2021). Introduction to the EQIPD quality system. ELife, 10, e63294.

[5] Bryans, J. S., Kettleborough, C. A., & Solari, R. (2019). Are academic drug discovery efforts receiving more recognition with declining industry efficiency? Expert Opinion on Drug Discovery, 14(7), 605–607.

[6] Patridge, E. V., Gareiss, P. C., Kinch, M. S., & Hoyer, D. W. (2015). An analysis of original research contributions toward FDA-approved drugs. Drug Discovery Today, 20(10), 1182–1187.

[7] Drenth, P. J. D. (2010). A European Code of Conduct for Research Integrity: (648332011-002) [Data set]. American Psychological Association.

[8] Voelkl, B., Vogt, L., Sena, E. S., & Würbel, H. (2018). Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLOS Biology, 16(2), e2003693. https://doi.org/10.1371/journal.pbio.2003693

7

General discussion

María Arroyo-Araujo

Groningen Institute for Evolutionary Life Sciences (GELIFES), University of Groningen (Groningen, The Netherlands)

SUMMARY OF RESULTS

The inconsistency of results within the preclinical field has raised concerns about the value of preclinical animal studies. Different approaches have been taken to unravel the possible source(s) of this discrepancies exemplified in Appendix 2. For example, the multi-site studies in **Chapter 2 & 3** showed how different degrees of alignment of experimental protocols between laboratories can influence the between-lab replicability of results. Briefly, we showed that harmonization of standardized protocols across laboratories can result in more consistent results than non-harmonization of protocols. Moreover, these chapters also indicate that there is room to optimize experimental protocols to not only decrease the between-lab variability of data, but also generate more robust results. For example, introducing systematic variation can improve the robustness of results across different contexts, which increases how likely they are to generalize and therefore enhances the value of preclinical studies. Accordingly, **Chapter 4** summarizes the variability of results across behavioral phenotyping studies using the Fmr1-KO mouse model by means of a systematic review and meta-analysis. The heterogeneity of results across studies could not be evaluated as a consequence of the diverse environmental conditions across experiments given the limited reporting in most studies; which underscores the importance of in detail reporting of methods and procedures to improve the reproducibility of studies. Despite the diverse experimental conditions across studies, 3 out of 16 phenotypes evaluated showed consistent effects across studies and aligned to the clinical profile; this highlights the importance of exploring the influence of environmental factors in phenotype expression in light of the robustness of results. Given the uncertainty of the influence of environmental factors in the heterogeneity of the aforementioned results, we decided to conduct a behavioral study using the Fmr1-KO mouse line to test whether the time of testing and/or the light intensity during testing could influence the phenotype expression. As described in **Chapter 5,** most of the behavioral outcomes were unaffected by the chosen environmental factors although the time of testing resulted in an enhanced response for the sensory processing task (*i.e.,* Acoustic Startle Response, ASR). Additionally, results indicated the order in which the tests were conducted altered the locomotion phenotype, which seemed to be enhanced in the KO animals only in the first test conducted and when the light brightness during test was set to dim. These results suggest that environmental factors other than the ones we included in our study may cause the large heterogeneity seen in the meta-analysis results. However, this remains to be confirmed. Finally, in **Chapter 6** and related **Appendix 1,** we aimed to address research improvements from a perspective beyond the research laboratory setting (*i.e.,* before and after conducting an experiment). These chapters present the EQIPD Quality System (QS) and its implementation in an academic lab setting to boost the quality of research data through the promotion of

effective planning, conducting and reporting of animal studies. Altogether, the work in this thesis underscores the importance of research practices that improve the quality of fundamental research in all the steps throughout the research process in such a way that the process will become more transparent and results will be more robust, generalizable and therefore more informative for both fundamental and translational research.

GENERAL DISCUSSION

Conflicting reports of preclinical findings are increasingly raising concerns regarding the reproducibility and replicability of preclinical results and the challenges that this represents for translational research. A method that allows to summarize scientific results while getting an overview of the consistency of results is to do a systematic reviews and meta-analysis. As exemplified in **Chapter 4,** there are big gaps where experimental details are missing in the published literature. This prevents studies from being reproduced while it also restricts the reader to critically judge the soundness of the experimental design and results interpretation. Moreover, results with incomplete experimental reporting may lose prediction power when incorporated to meta-analysis. Therefore, measures to improve reporting in research papers should be encouraged. For example, the risk of bias assessment adopted from clinical studies has served as a way to review the reporting of experimental design aspects relevant for the internal validity of the study such as: randomization, blinding, appropriate choice of outcome measure according to the goal of the study, complete outcome reporting. These research practices secure the internal validity of a study by assuring the causal relationship of the experimental manipulation and the outcome, while preventing biases such as selective reporting, selection bias, etc. Without the reporting of these practices it becomes unfeasible to critically judge the relevance of the results (1). In light of this, different guidelines such as the ARRIVE and PREPARE guidelines were developed to promote appropriate planning and reporting of animal studies (2,3). There are two studies that evaluated the compliance with quality checklists such as ARRIVE for animal studies submitted to PLOS ONE and Nature Publication Group. These studies showed that when compliance was only requested, no manuscript achieved full compliance (4); however, when compliance was mandatory, there was an increase of 16% in the reporting of randomization, blinding, sample size calculation and exclusions (5). These findings suggest that the adoption of such guidelines should be further promoted by universities, publishers, and funders to set an optimal standard on the reporting of preclinical studies (6), although interventions earlier on the research process might be more effective (*e.g.,* training student with such responsible research standards). Likewise, it has been recently reported that preclinical systematic review reports are often suboptimal, making it difficult to critically assess the

value of these summaries. Although the PRISMA guidelines for reporting of systematics reviews and metanalysis are often used, improving the reporting completeness, these do not fully address the transparency needed for the appraisal of preclinical systematic reviews (7).

Certainly, the adoption of transparent reporting practices would facilitate the reproducibility of studies and thus, the comparison and interpretation of results across laboratories. Moreover, transparent reporting of materials and methods will increase the scientific value of the scientific results since the influence that environmental variables could exert on the outcome measures can be formally assessed through systematic reviews and meta-analyses. Therefore, complete and transparent reporting of studies is an integral aspect for research improvement across scientific fields. Furthermore, scientific journals can promote good research practices as exemplified by the transparency and openness promotion (TOP) guidelines. These guidelines encourage journals to follow 8 standards regarding the research process (*i.e.,* reporting, pre-registration, data sharing etc.) including a process for exception for sharing for diverse reasons (*e.g.,* ethical concerns, intellectual property) all with the aim to promote transparency and openness to increase reproducibility and decrease research biases in particular selective reporting and publication bias (8).

The publication bias analysis presented in **chapter 4**, had the aim to explore the distribution of results around the overall effect captured in the meta-analysis. This allows one to judge whether reports may have been favored to be published because of the treatment effect they reported, meaning that studies with the opposite or null effect were not published (*i.e.*, publication bias). In chapter four, 4 out of the 14 behavioral categories showed publication bias. A way to decrease publication bias would be to pre-register animal studies (9). Pre-registration systems (*e.g.,* preclinicaltrials.eu) allow researchers to register their study plans in advance of the study; this ensures the transparent reporting of materials and methods (experimental and statistical) and the separation between exploratory and confirmatory research. The latter is important since it allows the reader to ponder the study accordingly. Moreover, as methods and analysis plans are publicly available, it allows for other researchers to give feedback or advise about potential improvements or setbacks. Worth mentioning, the pre-registered protocol can always be updated on the fly; for instance, after unexpected results. All the versions of the study protocol will be saved, justified and stay publicly available. Moreover, studies that need to be blinded for concerns of intellectual property or to protect a patent can pre-register under an embargo policy which makes the study private until there is no further risk. Even though when pre-registration has the potential to boost transparency of studies and prevent questionable research practices (results-

based analysis, p-hacking, selective outcome reporting), it should be further supported and promoted at a higher institutional level. For example, if funders where to require pre-registration to researchers that will run a confirmatory study of a pharmacological effect, pre-registering the study would promote complete and transparent reporting, appropriate study design, and responsible research practices; in addition, it may also reduce the publication pressure held by many researchers who have to obtain a minimum number of publications per year; preventing them from leaving aside the quality and completeness of the studies. In this way, implementing pre-registration of animal studies could improve the overall quality of the research while reducing the number of animals used as this practice would also aid to prevent the duplication of research studies. Funders could also ask to verify in pre-registration platforms whether the study to be funded has been (partially) performed or not already. The implementation of policies like pre-registration can be further supported by other strategies such as the use of quality systems like the one developed by EQIPD and included in this thesis **(Appendix 1)** (10). The use of such platforms would support rigorous research practices that go beyond the research laboratory setting that are likely to promote the generation of reliable preclinical data by assuring data is fit for the purpose intended (10). It would be interesting to measure the impact of implementing such a quality system in a lab overtime by, for example, looking at whether the published articles by the lab fulfill transparency and openness criteria more in-depth than before the quality system was implemented. It would be expected that after the quality system has been implemented, PhD students and involved staff would be more familiarized with open science practices and related responsible research practices, and therefore they would be more prone to use such resources. A similar intervention study could be planned to assess the degree of agreement of published papers with their pre-registered protocol compared to similar study plans for publicly funded projects alignment with the publication coming out of the project where the study plan. Nevertheless, the aforementioned strategies should not be seen as another rule to comply with, but as a change of the current mindset of the scientific community towards research improvement through the consistent use of responsible research practices.

Similarly to meta-research, multi-laboratory studies serve to evaluate the replication of results across laboratories while preserving relatively small sample sizes per laboratory but large sample size overall, increasing the predictive power of the study; they have also been proposed as generalizability studies (11). **Chapter 2 & 3** of this thesis explored the influence that the alignment of experimental designs across different laboratory sites can have on results replicability. In this sense, it has been noted that that rigorous standardization of study populations and conditions was failing to produce consistent results across different labs. In more recent years, this issue has been

addressed by exploring experimental designs that aim to incorporate the unavoidable variability of contexts between laboratories within a single laboratory (*e.g.,* diversified genetic composition of study populations and environmental conditions, multi-batch experiments) (12–15). These approaches are suggested to broaden the representativeness of the study population so results are more robust and likely generalizable and thus, more informative overall (16). Additionally, this approach raises awareness regarding the so common limited representation of the target study populations; specially, in biomedical research (*e.g.,* conducting clinical trials only in male participants while the target population includes females) both in the preclinical and clinical fields. Science as a system to advance knowledge should try to incorporate the diversity found in society whenever the aim is to benefit all its individuals and not only specific groups. Interestingly, introducing controlled systematic variation is also implemented in fields such as ecology/agronomy (13) where it has shown promising results; nevertheless and as presented in **chapter 5,** there is more research needed to detect those environmental factors that are most effective for introducing variation within studies without creating overly complex experimental design (and analysis), and that are the most feasible to diversify across labs and institutions. Even though, the multi-site studies presented in this thesis showed valuable results, multi-center studies are costly and although they offer advantages in terms of samples sizes over single-lab studies, they can also give a false sense of replicability given the thorough alignment of methods and protocols used across labs. This could result in less variable outcomes than those seen when taking a random sample of independent single-lab studies (17) thus, results should be critically appraised.

REPLICABILITY AS AN INDICATOR

As mentioned previously, the 'replicability crisis' in preclinical studies is being addressed with different approaches which aim to boost the quality of research where results replicability has been suggested as an indicator of unreliable results. However, it is known that preclinical research is not the only scientific field facing conflicting results between studies; this is also the case for psychology, economics, artificial intelligence and ecology (13,18,19). It has been reported that replicability rates of results in psychology, cancer biology, experimental economics and social and behavioral sciences range between 25 – 75% (18), indicating that the 'crisis' is spread throughout diverse scientific fields. However, this crisis can only be considered as such if we take replicability as an epistemic criterion that classifies findings as reliable or not. This strict dichotomy goes back to Karl Popper and although replicability has indeed served as a marker for 'good science' it fails to acknowledge the diversity of research questions, approaches and

entities studied across science (18,19). Taking these into account, one could realize that reproducibility and replicability are feasible and informative in those fields where there can be strict experimental control and the entity being measure is rather static over time and across contexts, for instance, physics or computer sciences. However, when approaching research fields with living systems in changing environments (e.g., laboratory animals, plants, social studies) we need to take into account the plasticity and historicity of the organism since it will play a big role when trying to replicate results with context sensitive organisms (14,18,20). Acknowledging the diversity of science and thus, the perks and limits of results replication can replace the crisis narrative with a mindset shift where researchers prioritize critical evaluation of the way studies are planned and conducted. One way to facilitate this may be as doing what researchers that cannot secure replicability do: thorough documentation of their data production process (*e.g.,* transparent reporting practices, open science, pre-registration of studies) -high reproducibility studies (19).

Specifically in preclinical studies, replicability would be most valuable as a strategy to interrogate the sources of outcome variation; irreplicable results in animal studies can indicate the limited inference space of results obtained presumably with sub-optimal research practices (*e.g.,* rigorous standardization of the study population, high risk of bias, questionable research practices). Nevertheless, there could be findings that are replicable but are still not informative because of the way the data was acquired or interpreted. Therefore, replicability in preclinical studies would be more useful as a warning sign for critical examination of scientific findings given their validity and generalizability value. Once replicability is used to address issues related to the study internal and external validity, researchers will be able to focus on how to the increase the robustness of results in light of their generalizability.

STUDY LIMITATIONS AND FUTURE PERSPECTIVES

A limitation of this thesis is that it did not cover all aspects of research improvement as research integrity or incentives within the scientific system; however, I hope it achieved its goal to raise awareness of the urgent need to steer the current research culture towards a more transparent, constructive and incentive oriented system. As much as responsible research practices support research improvement, it is necessary that institutions, funders and publishers also join sharing responsibility for enhancing research; for example, with incentives that promote a change in the research quality culture. In line with this, there are different aspects within the research culture that have room for improvement; these vary from smaller scale interventions such as funders introducing

guidelines for reproducibility and replicability activities into their merit-review criteria (6), or the training of PhD supervisors in responsible research practices (RRPs) to foster research integrity (21) up to larger scale incentives such as the adoption of open science (OS). Certainly, open science is a promising initiative with the potential to promote equity in the access to scientific knowledge by researchers from different institutions around the world that may not have the resources to cover for the expensive scientific journal subscriptions. Furthermore, it promotes data and methods transparency as all materials are open to the public, which in turn can result in the reuse of research data; this can also benefit researchers to generate 'low-cost' research while it would maximize the value of a single dataset (*i.e.,* more efficient research). Last but not least, open science could increase the likelihood of detecting research misconduct (8,22).

Similarly to OS, there are various actions that if implemented within the current scientific system that could support change towards a more transparent, inclusive and equitable research culture. For example, the current criteria for assessment of research careers are commonly focused on number of publications and citations and impact factor; however, it has been suggested that these measures are not optimal to evaluate research integrity and quality, plus they could be incentivizing suboptimal research practices. Certainly, the journal impact factor was originally developed for librarians to identify which journals to purchase; it is determined by technicalities that are unrelate to que scientific quality of the papers (23–25). Thus, assessing research careers with indicators such as the impact factor is rather uninformative regarding the trustworthiness, rigor and transparency of the research conducted. In contrast, there are diverse initiatives like the San Francisco Declaration on Research Assessment principles (DORA) (24), the Leiden Manifesto for research metrics (26) and the Hong Kong Principles for assessing researcher (27) that explore more suitable criteria to assess the quality of the research performed by, for example, looking into the responsible research practices implemented (*e.g.,* accurate and transparent reporting, the use of open science, pre-registration), acknowledge the broad range of research activities such as meta-research and validation studies, reward other research activities such as peer review and mentoring, to mention few. Just a broad and thorough assessment system is already being used in The Netherlands under the "Room for everyone's talents" (28). This program aims to acknowledge and reward diverse scientific paths that bibliometric indicators oversee. For example, time spent in education and patient care are activities that contribute to the output of scientific field and the current assessment system does not take into account; *i.e.,* assessing scientific careers by the number of publications seems unfair to those researchers that invest significant amount of their time in education or patient care and therefore are likely to have less publications than the researchers that are full-time doing research. In this way, it not only feasible to give equal importance to the different activities that are the back

bone of science but it also promotes a cultural shift towards a balance of individual and team-based research (29). Fostering cooperation in research creates a more inclusive work culture where diverse talents and expertise can interplay.

Equally important as implementing a better suite assessment for scientific careers is to support and empowers those who are starting their scientific career. Early career researchers (ECRs), defined as graduate/medical students, postdoctoral fellows and recently appointed independent researchers, represent the majority of the scientific workforce and the future leaders in science; yet, they are not often involved in decision-making roles and are rarely rewarded or incentivized for taking part in science improvement activities. This undermines and disincentives their science improvement efforts and prevents them from implementing changes to improve the scientific system. If the scientific community is striving to improve science for all, stakeholders should support science improvement efforts of skilled ECRs by involving them in decision-making processes, granting them funding and protected time for research improvement activities, empowering minorities and marginalized ECRs, and amplifying their research improvement efforts. By doing this, stakeholders would be also promoting equity, diversity and inclusion (EDI) in science as ECRs are a far more diverse cohort than senior scientists (age, genre, sexual identity, background, nationality, etc.). Supporting ECRs in science improvement and across scientific fields is a way to establish a culture of inclusion and equity to prevent bias and discrimination (30).

GENERAL CONCLUSION

Reproducibility and replicability of results in preclinical studies is suggested as a means to foster high quality research and thus robust and generalizable outcomes. In order to enhance research quality, there are numerous strategies that can be further implemented; however, it is necessary that researchers, institutions, funders, and publishers engage in changing the research culture. It is necessary to invest resources in training researchers and students in researcher(er) integrity topics and the most common causes of irreproducible and irreplicable results (e.g., statistical misuse, inappropriate experimental design, misconduct) (31). Additionally, changing the research reward and incentive system would further facilitate the change in research culture if funding and publication do not prioritize outstanding results over robust and meaningful results. A change in research culture should include an EDI perspective as a way to promote equity in society.

In practice, aiming to produce results that can be replicated in other labs would entail but not be limited to 1) pre-register the study plan to ensure the study protocol complies with quality standards like blinding and randomization, and transparent reporting of the aim, methods and analysis of the study (to ensure reproducibility). If pre-registration is not an option, following guidelines such as PREPARE and ARRIVE is highly recommended. 2) critically assessment on whether the experimental design is fit for the purpose of the study, for instance, does the population sample truly represent the target population (*i.e.,* gene-by-environment interactions). 3) conduct the experiment and analysis following research integrity principles according to the study plan. 4) report all data outcomes and share the results, data/code in an open access platform. 5) publish in an open access journal whenever possible.

REFERENCES

1. Macleod M, Mohan S. Reproducibility and Rigor in Animal-Based Research. ILAR Journal. 2019 Dec 31;60(1):17–23.
2. Sert NP du, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, et al. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLOS Biology. 2020 Jul 14;18(7):e3000411.
3. Smith AJ, Clutton RE, Lilley E, Hansen KEA, Brattelid T. PREPARE: guidelines for planning animal research and testing. Lab Anim. 2018 Apr 1;52(2):135–41.
4. Hair K, Macleod MR, Sena ES, Sena ES, Hair K, Macleod MR, et al. A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). Research Integrity and Peer Review. 2019 Jun 12;4(1):12.
5. Macleod MR, Group TNC. Findings of a retrospective, controlled cohort study of the impact of a change in Nature journals' editorial policy for life sciences research on the completeness of reporting study design and execution [Internet]. bioRxiv; 2017 [cited 2022 Jul 13]. p. 187245. Available from: https://www.biorxiv.org/content/10.1101/187245v1
6. National Academies of Sciences, Engineering, and Medicine. Reproducibility and Replicability in Science. [Internet]. Washington, DC: The National Academies Press.; 2019. Available from: https://doi.org/10.17226/25303.
7. Hunniford VT, Montroy J, Fergusson DA, Avey MT, Wever KE, McCann SK, et al. Epidemiology and reporting characteristics of preclinical systematic reviews. PLOS Biology. 2021 May 5;19(5):e3001177.
8. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015 Jun 26;348(6242):1422–5.
9. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. PNAS. 2018 Mar 13;115(11):2600–6.
10. Bespalov A, Bernard R, Gilis A, Gerlach B, Guillen J, Castagne V, et al. Introduction to the EQIPD quality system. Zaidi M, editor. eLife. 2021 May 24;10:e63294.
11. Mogil JS, Macleod MR. No publication without confirmation. Nature. 2017 Feb;542(7642):409–11.
12. Karp NA, Wilson Z, Stalker E, Mooney L, Lazic SE, Zhang B, et al. A multi-batch design to deliver robust estimates of efficacy and reduce animal use – a syngeneic tumour case study. Sci Rep. 2020 Apr 10;10(1):6178.
13. Milcu A, Puga-Freitas R, Ellison AM, Blouin M, Scheu S, Freschet GT, et al. Genotypic variability enhances the reproducibility of an ecological study. Nat Ecol Evol. 2018 Feb;2(2):279–87.
14. Voelkl B, Würbel H. A reaction norm perspective on reproducibility. Theory Biosci. 2021 Jun;140(2):169–76.
15. von Kortzfleisch VT, Karp NA, Palme R, Kaiser S, Sachser N, Richter SH. Improving reproducibility in animal research by splitting the study population into several 'mini-experiments.' Sci Rep. 2020 Oct 6;10(1):16579.
16. Usui T, Macleod MR, McCann SK, Senior AM, Nakagawa S. Meta-analysis of variation suggests that embracing variability improves both replicability and generalizability in preclinical research. PLOS Biology. 2021 May 19;19(5):e3001009.
17. Lewis M, Mathur MB, VanderWeele TJ, Frank MC. The puzzling relationship between multi-laboratory replications and meta-analyses of the published literature. R Soc Open Sci. 2022 Jan 10;9(2):211499.
18. Guttinger S. The limits of replicability. Euro Jnl Phil Sci. 2020 Jan 15;10(2):10.

19. Leonelli S. Re-Thinking Reproducibility as a Criterion for Research Quality [Internet]. 2018 [cited 2022 Jul 12]. Available from: http://philsci-archive.pitt.edu/14352/
20. Voelkl B, Vogt L, Sena ES, Würbel H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLOS Biology. 2018 Feb 22;16(2):e2003693.
21. Haven T, Bouter L, Mennen L, Tijdink J. Superb supervision: A pilot study on training supervisors to convey responsible research practices onto their PhD candidates. Accountability in Research. 2022 Apr 27;0(0):1–18.
22. Gopalakrishna G, Riet G ter, Vink G, Stoop I, Wicherts JM, Bouter LM. Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. PLOS ONE. 2022 Feb 16;17(2):e0263023.
23. Per O Seglen. Why the impact factor of journals should not be used for evaluating research. BMJ. 1997;314:498–502.
24. Read the Declaration [Internet]. DORA. [cited 2022 Oct 26]. Available from: https://sfdora.org/read/
25. Not-so-deep impact. Nature. 2005 Jun;435(7045):1003–4.
26. Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. Nature. 2015 Apr;520(7548):429–31.
27. Moher D, Bouter L, Kleinert S, Glasziou P, Sham MH, Barbour V, et al. The Hong Kong Principles for assessing researchers: Fostering research integrity. PLOS Biology. 2020 Jul 16;18(7):e3000737.
28. Recognition and rewards of academics [Internet]. Universities of the Netherlands. Available from: https://www.universiteitenvannederland.nl/en_GB/Recognition-and-rewards-of-academics.html
29. Quality in preclinical research: Networked rather than alone [Internet]. [cited 2022 Oct 22]. Available from: https://www.volkswagenstiftung.de/en/news-press/news/quality-in-preclinical-research-networked-rather-than-alone
30. Kent BA, Holman C, Amoako E, Antonietti A, Azam JM, Ballhausen H, et al. Recommendations for empowering early career researchers to improve research culture and practice. PLOS Biology. 2022 Jul 7;20(7):e3001680.
31. Macleod M, the University of Edinburgh Research Strategy Group. Improving the reproducibility and integrity of research: what can different stakeholders contribute? BMC Research Notes. 2022 Apr 25;15(1):146.


A1

Introduction to the EQIPD Quality System

Anton Bespalov[1*], René Bernard[2*], Anja Gilis[3*], Björn Gerlach[1*], Javier Guillen[4], Vincent Castagné[5], Isabel A. Lefevre[6], Fiona Ducrey[7], Lee Monk[8], Sandrine Bongiovanni[9], Bruce Altevogt[10], **María Arroyo-Araujo**[11], Lior Bikovski[12,13], Natasja de Bruin[14], Esmeralda Castaños-Vélez[2], Alexander Dityatev[15,16,17], Christoph H. Emmerich[1], Raafat Fares[18], Chantelle Ferland-Beckham[19], Christelle Froger-Colléaux[5], Valerie Gailus-Durner[20], Sabine M. Hölter[21], Martine Hofmann[14], Patricia Kabitzke[22,23], Martien J. Kas[11], Claudia Kurreck[2], Paul Moser[24,25], Malgorzata Pietraszek[1], Piotr Popik[26], Heidrun Potschka[27], Ernesto Prado Montes de Oca[28,29,30], Leonardo Restivo[31], Gernot Riedel[32], Merel Ritskes-Hoitinga[33,34], Janko Samardzic[35], Michael Schunn[36], Claudia Stöger[20], Vootele Voikar[37], Jan Vollert[38], Kim Wever[33], Kathleen Wuyts[39], Malcolm Macleod[40], Ulrich Dirnagl[2,41], Thomas Steckler[3]

* these authors contributed equally to this work

[1]PAASP, Heidelberg, Germany; [2]Department of Experimental Neurology, Charité Universitätsmedizin, Berlin, Germany; [3]Janssen Pharmaceutica NV, Beerse, Belgium; [4]AALAC International, Pamplona, Spain; [5]Porsolt, Le Genest-Saint-Isle, France; [6]Rare and Neurologic Diseases Research, Sanofi, Chilly-Mazarin, France; [7]Integrity and Global Research Practices, Sanofi, Chilly-Mazarin, France; [8]Research and Clinical Development Quality, UCB, Slough, UK; [9]Quality Assurance, Novartis Institutes for BioMedical Research, Novartis Pharma, Basel, Switzerland; [10]Pfizer, Silver Spring, MD, USA; [11]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands; [12]The Myers Neuro-Behavioral Core Facility, Sackler School of Medicine, Tel Aviv University, Israel; [13]School of Behavioral Sciences, Netanya Academic College, Netanya, Israel; [14]Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Branch for Translational Medicine and Pharmacology TMP, Frankfurt am Main, Germany; [15]Molecular Neuroplasticity, German Center for Neurodegenerative Diseases, Magdeburg, Germany; [16]Center for Behavioral Brain Sciences, Magdeburg, Germany; [17]Medical Faculty, Otto-von-Guericke University, Magdeburg, Germany; [18]Charles River Laboratories, Safety Assessment, Lyon, France; Current affiliation: Etisense SAS, R&D department, Lyon, France; [19]Cohen Veterans Bioscience, Boston, MA, USA; [20]German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [21]Institute of Developmental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, and Technical University Munich, Germany; [22]PAASP US, Ridgefield, CT, USA; [23]The Stanley Center for Psychiatric Research, Broad Institute of MIT & Harvard, Cambridge, MA 02142; [24]Cerbascience, Toulouse, France; [25]PAASP France, Toulouse, France; [26]Maj Institute of Pharmacology, Polish Academy of Sciences, Krakow, Poland; [27]Institute of Pharmacology, Toxicology and Pharmacy, Ludwig-Maximilians-University, Munich, Germany; [28]Research Center in Technology and Design Assistance of Jalisco State, National Council of Science and Technology, Guadalajara, Jalisco, Mexico; [29]Scripps Research Translational Institute, La Jolla, CA, USA; [30]Integrative Structural and Computational Biology, Scripps Research, La Jolla, CA, USA; [31]Neuro-BAU, Department of Fundamental Neurosciences, University of Lausanne, Lausanne, Switzerland; [32]Institute of Medical Sciences, University of Aberdeen, Scotland, UK; [33]SYRCLE, Department for Health Evidence, Radboud University Medical Center, Nijmegen, the Netherlands; [34]Department for Clinical Medicine, Aarhus University, Denmark; [35]Institute of Pharmacology, Medical Faculty, University of Belgrade, Serbia; [36]Institute of Science and Technology, Klosterneuburg, Austria; [37]Neuroscience Center and Laboratory Animal Center, Helsinki Institute of Life Science, University of Helsinki, Finland; [38]Pain Reearch, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK; [39]Avertim, Brussels, Belgium; [40]Centre for Clinical Brain Sciences, University of Edinburgh, Scotland, UK; [41]QUEST Center for Transforming Biomedical Research, Berlin Institute of Health, Germany

Published in eLife 2021 (10:e63294.)

ABSTRACT

While high risk of failure is an inherent part of developing innovative therapies, it can be reduced by adherence to evidence-based rigorous research practices. Numerous analyses conducted to date have clearly identified measures that need to be taken to improve research rigor. Supported through the European Union's Innovative Medicines Initiative, the EQIPD consortium has developed a novel preclinical research quality system that can be applied in both public and private sectors and is free for anyone to use. The EQIPD Quality System was designed to be suited to boost innovation by ensuring the generation of robust and reliable preclinical data while being lean, effective and not becoming a burden that could negatively impact the freedom to explore scientific questions. EQIPD defines research quality as the extent to which research data are fit for their intended use. Fitness, in this context, is defined by the stakeholders, who are the scientists directly involved in the research, but also their funders, sponsors, publishers, research tool manufacturers and collaboration partners such as peers in a multi-site research project. The essence of the EQIPD Quality System is the set of 18 core requirements that can be addressed flexibly, according to user-specific needs and following a user-defined trajectory. The EQIPD Quality System proposes guidance on expectations for quality-related measures, defines criteria for adequate processes (i.e., performance standards) and provides examples of how such measures can be developed and implemented. However, it does not prescribe any pre-determined solutions. EQIPD has also developed tools (for optional use) to support users in implementing the system and assessment services for those research units that successfully implement the quality system and would like to seek formal accreditation. Building upon the feedback from users and continuous improvement, a sustainable EQIPD Quality System will ultimately serve the entire community of scientists conducting non-regulated preclinical research, by helping them generate reliable data that are fit for their intended use.

THE CHALLENGE: DISCOVERY OF NOVEL THERAPIES REQUIRES RIGOR IN RESEARCH PRACTICES

The success rate in the discovery of novel, safe and effective pharmacotherapies has been declining steadily over the last few decades (Scannell et al., 2012). There are several factors likely accounting for this unfortunate record (DiMasi et al., 2016; Waring et al., 2015; Shih et al., 2018). While some of these factors (e.g., deeper knowledge of disease biology or clinical trial methodology) will take years, if not decades, of continued research to be properly addressed, others can be readily controlled today (Bespalov et al., 2016; Landis et al., 2012). One area requiring immediate attention is research rigor, which is estimated to be lacking in 50-90% of preclinical studies (Freedman et al., 2012).

High risk of failure is an inherent part of developing innovative therapies (DiMasi et al., 2016). However, some risks can be greatly reduced and avoided by adherence to evidence-based rigorous research practices. Indeed, numerous analyses conducted to date have clearly identified measures that need to be taken to improve research rigor (Begley and Ioannidis, 2015; Landis et al., 2012; Ritskes-Hoitinga and Wever, 2018; Vollert et al., 2020; Volsen and Masson, 2009).

THE EQIPD CONSORTIUM: ENHANCING RESEARCH QUALITY AS THE MAIN OBJECTIVE

Improving research rigor has biomedical, societal, personal, economic and ethical benefits for academia and industry alike, since the development of novel therapies is often rooted in academic discoveries and requires a highly specialized effort of industry to translate these discoveries into clinically useful applications. Moreover, this simple dichotomy between purely academic research and large industry/big pharma efforts is currently being replaced by networks of biotechs, spin-offs, private and public funders, contract research organizations (CROs), academic institutions engaging in drug discovery projects and manufacturers of research tools. It is therefore important that strategies to increase the robustness and reliability of preclinical research, both in terms of conduct and reporting, involve all these different stakeholders.

To address this challenge in preclinical biomedical research in a collaborative manner, the Enhancing Quality in Preclinical Data (EQIPD; originally called European Quality in Preclinical Data) consortium was formed in 2017 with founding members from 29 institutions across 8 different countries (https://quality-preclinical-data.eu). The consortium works closely with a large group of associated collaborators, advisors and stakeholders

representing research institutions, publishers, funders, learned societies and professional societies, from nearly 100 organizations in Europe and the US.

Supported through the European Union's Innovative Medicines Initiative (IMI), the EQIPD consortium, among other deliverables, aimed to develop a novel preclinical research quality system that can be applied in both the public and private sectors. Such a quality system should be suited to boost innovation by ensuring the generation of robust and reliable preclinical data while being lean, effective and not becoming a burden that could negatively impact the freedom to explore scientific questions.

EQIPD defines research quality as the extent to which research data are fit for intended use (for related definitions and explanations, see Juran and Godfrey, 1999; Gilis, 2020). Fitness, in this context, is defined by the stakeholders, who can be scientists themselves, but also patients, funders, sponsors, publishers and collaboration partners (e.g., peers in a multi-site research project).

The EQIPD consortium has developed a quality system that is free for anyone to use. Further, EQIPD is preparing training support and assessment services for those research units that successfully implement the quality system and would like to seek formal accreditation.

A NEW QUALITY SYSTEM TO BOOST INNOVATION

Quality systems usually appear as a response to an existing need (Table 1). For example, the development of the Good Laboratory Practice (GLP) standards, introduced first by the Food and Drug Administration (FDA) in the late 1970s, was triggered by poor research practices that compromised human health, such as mis-identification of control and experimental animals, omitted, non-reported or suppressed scientific findings, data inventions, dead animal replacements and mis-dosing of test animals (Bongiovanni et al., 2020; Marshall, 1983). In the Organisation for Economic Co-operation and Development (OECD) Principles (https://www.oecd.org/chemicalsafety/testing/overview-of-good-laboratory-practice.htm), GLP is defined as "a quality system concerned with *the organisational process and the conditions* under which non-clinical health and environmental safety studies are planned, performed, monitored, recorded, archived and reported".
GLP is a standard approach to quality in the regulated areas of preclinical drug development (which largely relate to non-clinical safety and toxicology studies rather than efficacy; see Supplement S1 Glossary for a definition of regulated research), where trained

TABLE 1 --- Comparison of quality systems

Quality system	ISO 9001	GLP (FDA, OECD)	EQIPD
Year Launched	1987, 2015	1976, 1981	2020
Application area	A general QMS that can be applied to all aspects of organizations (not focused on biomedical research)	Non-clinical health and environmental safety studies upon which hazard assessments are based	Non-regulated preclinical (non-clinical) biomedical research
Initial stimulus to be developed	Procuring organizations needed a basis of contractual arrangements with their suppliers (i.e., basic requirements for a supplier to assure product quality)	Regulators such as FDA aimed to avoid poorly managed or fraudulent non-clinical studies on safety of new drugs	Biomedical research community (industry and academia) recognized the negative impact of lacking research rigor on the development of novel therapeutics, and the need for a comprehensive practical solution to help enhance preclinical data reliability
Objectives	To certify that a product (which can be preclinical data) or a service is provided with consistent, good-quality characteristics, which satisfy the stated or implied needs of customers	To ensure the quality, integrity and reliability of data on the properties and/or safety of test items concerning human health and/or the environment	To facilitate generating robust and reliable preclinical data and thereby boost innovation
Customers	Typically outside of the organization (anyone who requires a product or service)	Typically outside of the organization (patients, regulators, sponsors, etc.)	In most cases, both inside (scientists themselves) and outside (patients, funders, collaboration partners, publishers, etc.) of the organization
Main focus	Standardization of processes The organizational overall performance is continuously improved (process approach) to enhance customer satisfaction and development initiatives are done on a sound basis for sustainability	The organizational process and the conditions under which non-clinical health and environmental safety studies are planned, performed, monitored, recorded, archived and reported	The outcome of research activities that is robust, reliable, traceable, properly recorded, reconstructible, securely stored and trustworthy (generated under appropriately unbiased conditions)
Dedicated quality professionals	Not required (advisable for larger organizations)	Required	Not required (advisable for larger organizations)
Formal training on implementation and use	Not required	Required	Advisable, but not required
Assessments	External (ISO auditors) and internal (internal auditors)	External (health authorities / governmental inspectors) and internal (QA auditors)	Self-assessment (by Process Owner), external (by EQIPD)[1]

[1] additional internal assessments may be conducted by qualified colleagues (e.g., dedicated quality professionals) outside the research unit but within the same organization (advisable for larger organizations)

personnel perform mainly routine analyses, following defined Standard Operating Procedures (SOPs), and deliver data ultimately supporting patient safety.

There have been attempts to develop a quality system based on GLP – i.e., taking GLP as the basis and eliminating elements that are seen as excessive for the purposes of non-regulated drug discovery. However, GLP does not provide explicit guidance regarding those aspects of study design, conduct, analysis and reporting that are important to minimize the risk of bias and make research robust. In other words, even if it were made less demanding, conventional GLP cannot address some of today's key challenges in non-regulated preclinical research.

In contrast, the EQIPD Quality System is a novel system specifically aimed at supporting innovation in preclinical biomedical research. While the direct consequence of installing a quality system will be the generation of research data that are of higher rigor, the ultimate goal is to improve the efficiency of developing novel effective and safe therapies.

DEVELOPMENT OF A NEW QUALITY SYSTEM BY EQIPD

EQIPD was started in October 2017 and during the first phase (until June 2018), three work packages of the EQIPD consortium have delivered:

- A systematic review of guidelines for internal validity in the design, conduct and analysis of research involving laboratory animals (Vollert et al., 2020);
- An inventory of current practices and expectations towards quality management in non-regulated preclinical research (based on interviews with 70 consortium members and stakeholders);
- A review and analysis of governance in existing quality management systems (AAALAC International; ASQ Best Quality Practices for Biomedical Research in Drug Development; BBSRC Joint Code of Practice; ISO 9001, ISO 17025, ISO 15189; Janssen discovery quality system; Novartis research quality system; OECD Principles of GLP; RQA – Quality Systems Workbook).

During the second phase (July 2018 - January 2019), a working group was assembled from the EQIPD consortium members (n=20). Based on the collected information, the working group nominated 75 statements that could define a functional quality system in non-regulated research. After three Delphi feedback rounds and two consensus meetings, these statements were revised, resulting in a final list of 18 core requirements (Table 2; see below for details).

During the third phase (February 2019 – September 2019), a supporting framework was developed (see below) and pilot implementation of the quality system started at four independent research sites.

TABLE 2 --- EQIPD Core Requirements

Categories	#	Item
Research team	1	Process Owner must be identified for the EQIPD Quality System
	2	Communication process must be in place
Quality culture	3	The research unit must have defined quality objectives and rules to reach them
	4	All activities must comply with relevant legislation and policies
	5	The research unit must have a procedure to act upon concerns of potential misconduct
Data integrity	6	Generation, handling and changes to data records must be documented
	7	Data storage must be secured at least for as long as required by legal, contractual or other obligations or business needs
	8	Reported research outcomes must be traceable to experimental data
	9	Reported data must disclose all repetitions of a study, an experiment, or a test regardless of the outcome
	10	Investigator must declare in advance whether a study is intended to inform a formal knowledge claim
Research processes	11	All personnel involved in research must have adequate training and competence to perform assigned tasks
	12	Protocols for experimental methods must be available
	13	Adequate handling and storage of samples and materials must be ensured
	14	Research equipment and tools must be suitable for intended use and ensure data integrity
Continuous improvement	15	Risk assessment must be performed to identify factors affecting the generation, processing and reporting of research data
	16	Critical incidents and errors during study conduct must be analyzed and appropriately managed
	17	An approach must be in place to monitor the performance of the EQIPD Quality System, and address identified issues
Sustainability	18	Resources for sustaining the EQIPD Quality System must be available

Based on the feedback from those pilot implementation sites and interactions with the stakeholder group, an updated version of the framework was released for beta-testing in November 2019. The final version of the quality system was released in September 2020.

THE EQIPD QUALITY SYSTEM: KEY FEATURES

Table 3 presents five principles on which the EQIPD Quality System is based.

These principles communicate in a maximally concise and direct form that EQIPD Quality System supports scientists in triggering changes in research practices, helps to identify objectives and direction of change but does not prescribe any specific solutions as long as the research processes are kept transparent and traceable.

TABLE 3 --- Key principles

Principle	Explanation	Example
Engage with autonomy	Decisions about specific needs and solutions are made by researchers, and not by EQIPD. EQIPD has formulated core requirements for the QS implementation and, as a partner in this process, EQIPD asks critical questions and provides recommendations that are voluntary to follow and are provided only to help the researchers throughout the implementation and use.	EQIPD recommends applying randomization to all studies but it is up to the researcher to decide whether randomization is applying to a particular study or a particular study design
Grow through reflection	What it means to have the right quality level in place is suggested by your environment (collaborators, funders, institution, etc.). EQIPD does not "invent" needs or requirements of your funders or your collaborators. As a partner in this process, EQIPD QS only allows you to see these requirements better and suggests ways of implementing them (Gilis, 2020).	EQIPD identifies overlapping requirements from different stakeholders towards the use and reporting of randomization.
Focus on goal	Focus on the outcome (performance standards), not on the path, timelines or the tools to get there (Guillen, 2010).	EQIPD highlights the importance of "randomness" (lack of pattern or predictability) in the correctly developed randomization sequence but leaves it up to the user to select a specific method or tool.
Be transparent	Key research processes must be transparent. This principle applies specifically to retention and accessibility of information related to key decisions related to study design, conduct or analysis (e.g., decisions to include or exclude certain data points in the analysis).	If you decide not to apply randomization, the decision must be stated and must be justified, recorded and reported.
Leave a trace	Key research processes must be traceable. Complementary to the principle above, this principle refers to retention and accessibility of all information that is necessary for a complete reconstruction of a key research process (e.g., raw data related to reported data are findable, and reported data are reconstructable from raw data).	If you do apply randomization, the way you apply randomization must be traceable and reported

The EQIPD Quality System will deal with highly diverse research environments and associated challenges. The five principles are, therefore, instrumental in finding answers to specific questions – e.g., is this particular practice in line with the EQIPD expectations? or should this particular process be documented?

FLEXIBLE: DRIVEN BY THE NEEDS OF AN INDIVIDUAL RESEARCH UNIT

Research environments are highly diverse: the needs of researchers at a big pharma company are different from those at a biotech; the needs of CROs are different from those of academic labs, etc. Thus, improving data quality is a challenge that cannot be tackled using a one-size-fits-all solution and flexibility is a critical requirement for future success.

The EQIPD Quality System is flexible: researchers are not confronted with a long and ultimate A-to-Z list of what should be done and in what sequence. Instead, implementation of the EQIPD Quality System is characterized by:

- user-specific content – i.e., the exact nature of the individual elements of the EQIPD Quality System are defined largely by the users and their environment;
- a variable trajectory – i.e., there are very limited expectations regarding the sequence of introducing the different elements of the EQIPD Quality System; and
- no deadlines or fixed timelines – i.e., each unit adopts the EQIPD Quality System at its own pace, depending on the existing needs and available resources.

EQIPD has developed tools (for optional use) that help scientists identify and organize information to address their own customized needs (e.g., related to *my* research funding source, *my* national regulations for the use of animals, expectations of *my* collaboration partners, policies set by *my* institution, *my* own commitment to research rigor, etc.). Being unique to a research unit or a researcher, such needs can be very specific to local or personal circumstances (i.e., essential for *my* success, *my* funding, *my* career, for instance because of the requirements of *my* preferred funder), and as such may be addressed with a higher or lower priority. Based on these factors, each research unit or researcher can determine their sequence of actions (Figure 1). EQIPD tools offer examples and ready-to-use solutions as well as information to develop new user-specific solutions.

FIGURE 1

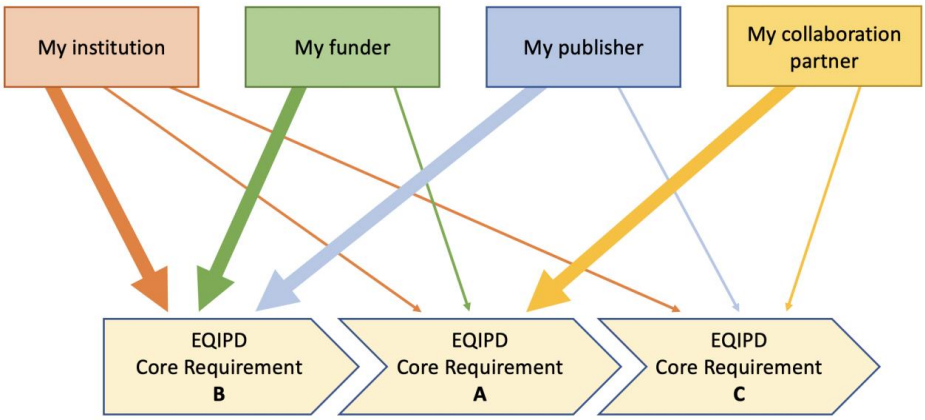


Figure 1: Flexible sequence of implementation of the EQIPD core requirements. Depending on the current needs, a research unit may prioritize the implementation of one or another core requirement. For example, tasks related to core requirement "B" are highly relevant for the research unit's parent institution, the funding organization and a scientific journal where the research team plans to publish the results of their work. In contrast, core requirement "C" is of lower importance and can, therefore, be addressed at a later timepoint.

For example, EQIPD has reviewed research quality expectations of several major public funders and pharmaceutical companies. Summaries of these expectations as well as examples of how these expectations can be met are available for downloading from the EQIPD's online Toolbox.

TEAM EFFORT: UNDERSTANDING AND ENDORSING RESEARCH QUALITY OBJECTIVES

The focus on the specific needs of an individual research unit is ensured by the Process Owner, a person within the organization who has access to the necessary resources, and the competence and the authority to implement all steps needed to establish the EQIPD Quality System. Typically, the Process Owner should be someone who directs the work of the research unit (e.g., group leader, principal investigator, CEO or department head) and is knowledgeable about the importance of quality in research. EQIPD expects the Process Owner to be identified at the very first step of implementing the EQIPD Quality System (Table 2; core requirement #1).

In the second step, the Process Owner defines the scope - i.e., the research unit (lab, territory, organization or part thereof) where the EQIPD Quality System will be applied - and identifies colleagues who will be actively involved in working on the implementation, as well as those who will be informed and may need to be trained about the new process (core requirement #2). To that end, the Process Owner sets up a communication plan to support the team's buy-in and to facilitate two-way information flow, in order to also capture feedback related to performance of the existing and newly introduced practices.

EQIPD also expects research units to define quality objectives (core requirement #3). Although it may sound formal, this core requirement is indispensable and should be articulated at a level understandable and meaningful to everyone in the research unit.

Why are quality objectives needed? Once the Process Owner has decided to accept the role and responsibilities and has defined the research unit where the EQIPD Quality System will be implemented, it is worth getting prepared to answer questions that will likely come from colleagues inside and outside of the research unit: why are we doing this if, at least today, no such quality system is required by funders or collaboration partners and if, at least on first sight, we can successfully meet the goals without changing anything?

The answer to these questions helps justify the efforts and time to be invested in the implementation and maintenance of the quality system. It also provides an argument by balancing the potentially negative impact on traditional metrics of scientific success (e.g., fewer positive results generated, more time needed to complete projects) against the value of higher quality research (greater confidence in the results and scientific interpretations when results are shared with peers or published, improved rigor in decision making, publication in high-profile journals, etc.).

In EQIPD terms, the answer should be documented as a mission statement, i.e., a concise summary of why quality matters for that specific research unit. EQIPD provides examples of how scientists working in different roles and at various types of organizations may answer the question "why quality matters" (https://osf.io/vduze).

It is important that the mission statement is understood, willingly accepted and followed by all members of the research unit.

If a Process Owner, alone or together with the research team members, cannot generate a clear and convincing answer to this question, no further steps should be taken and the implementation of the quality system is best postponed until a good answer is found and the research team is willing to accept a quality mindset.

EQIPD QUALITY SYSTEM AS PART OF THE OVERALL ORGANIZATIONAL QUALITY CULTURE

The Process Owner may also be asked and should be prepared to explain that the EQIPD Quality System does not replace and does not intend to re-interpret any of the existing rules, policies and other quality systems (which focus on specific areas) that apply to the research unit's environment.

EQIPD mandates that "all activities must comply with relevant legislation and policies" (core requirement #4) and that a "research unit must have a procedure to act upon concerns of potential misconduct" (core requirement #5). If, for the vast majority of organizations, no additional effort will be required to meet these expectations, why are they included in the list of core requirements?

First, EQIPD does not want to be associated with organizations that engage in or tolerate unacceptable ethical practices or legal violations.

Second, the EQIPD Quality System is focused on quality, not legislation. Legislation may differ from country to country and for different research activities; hence, it is not possible to specify these individually in the EQIPD Quality System. Furthermore, EQIPD cannot oversee the way an organization deals with the legal requirements of, e.g., handling hazardous substances, but emphasizes the need for compliance with such regulations as a basis on which all other quality measures rest.

Another example concerns the care and use of laboratory animals that play a pivotal role in the research process. Society has granted the biomedical research community with the privilege to use laboratory animals in research under very specific conditions, all aiming to prevent inappropriate use of these ethically highly sensitive resources. Clearly, it is not acceptable to waste animals due to poor study design, conduct or analysis.

Ethical concerns on the use of animals in research have promoted the creation of a legal framework in almost every country (e.g., Animal Welfare Act in the US; Directive 2010/63 in the EU). Scientific evidence demonstrates that many aspects of animal care and use that are beyond the legal requirements have a direct impact on research results (Guillén and Steckler, 2020). The EQIPD team has developed a concise checklist that allows scientists to review if their animal care and use processes meet at least a minimum standard that supports the implementation and maintenance of the EQIPD Quality System. This review could optionally serve as the basis for further, more specific accreditation of the animal care and use program (i.e., AAALAC International accreditation) to ensure the

implementation of high standards of animal care and use that would further contribute to increasing the quality of research (Supplement S2 Animal care and use checklist).

EQIPD-DEFINED PRINCIPLES, USER-DEFINED CONTENT

Implementation of the EQIPD Quality System does not require researchers to stop or reduce ongoing experimental work. It is designed so that it takes only minimal effort to sign up and begin the journey towards a quality system that should help researchers gradually improve certain quality aspects of their work.

The EQIPD Quality System gives guidance on expectations for quality-related measures, defines criteria for adequate processes (i.e., performance standards) and provides examples of how such measures can be developed and implemented. However, it does not prescribe any pre-determined solutions. Rather, users define their own specific solutions tailored to their individual settings.

For example, integrity of research data is one of the central concepts that the EQIPD Quality System aims to support. Four core requirements define the desired outcomes for raw data generation and handling (core requirement #6), data storage (core requirement #7), data traceability (core requirement #8), and transparency of reported data (core requirement #9). Thus, the "what" is clearly described. However, there are various ways to fulfil these requirements. For instance, secure data storage could be achieved by using conventional paper-based laboratory notebooks, electronic laboratory notebooks, custom-built electronic solutions or paper-based controlled-access archives. Thus, there is flexibility in how integrity of research data could be achieved, and it is for the users of the system to decide on the best solution for their specific situation.

FOCUSED ON THE GENERATION OF FIT-FOR-PURPOSE RESEARCH DATA

In general, EQIPD recommends that scientists apply protection against risks of bias for every study and unambiguously disclose the protective measures used. Each study has a particular purpose and the rigor applied to the study should be defined, documented in advance and be commensurate with the purpose of the study.

There are modes of research that can tolerate a certain level of uncertainty and do not lead to a formal knowledge claim (see Supplement S1 Glossary for definition). Such work

is an essential part of the research process and may be used to generate hypotheses or to provide evidence to give the investigator greater confidence that an emerging hypothesis is valid, to develop new methods or to "screen" compounds for potential effects prior to more formal testing.

There are also modes of research where researchers cannot accept inadequate control of the risks that can bias the research results (Dirnagl, 2016; Hooijmans et al., 2014). For research that is conducted with the prior intention of informing a knowledge claim, EQIPD requires that maximal possible rigor is applied (and exceptions explained and documented in the study plan; see Table 4). Such research will usually (but not always) involve some form of null hypothesis statistical testing or formal Bayesian analysis. Here, hypotheses are articulated in advance of data collection, with pre-specified criteria defining the primary outcome measure and the statistical test to be used.
Examples of research requiring maximal possible rigor may include:

TABLE 4 --- Expectations towards rigor in study design

	All research	Research informing a formal knowledge claim (i.e., research requiring maximal rigor)
Study plan	Should be defined and documented before starting the experiments	Must be defined and documented before starting the experiments
Study hypothesis	Advised to define	Must be pre-specified
Blinding	Advised to implement	Should be implemented, exceptions must be justified and documented
Randomization	Advised to implement	Should be implemented, exceptions must be justified and documented
Sample size calculation	Advised to define and document before starting the experiments	Must be defined and documented before starting the experiments (e.g., included in the study plan)
Data analysis	Advised to define and document before starting the experiments	Must be defined and documented before starting the experiments (e.g., as a formal statistical analysis plan and/or included in the study plan)
Inclusion and exclusion criteria	Advised to define and document before starting the experiments	Must be defined and documented before starting the experiments (e.g., included in the study plan)
Deviations from study plan	Advised to document	Must be documented
Preregistration	-	Should be implemented

- Experimental studies to scrutinize preclinical findings through replication of results (Kimmelman et al., 2014);
- Research aimed at generating evidence that enables decisions which will invoke substantial future investment (e.g., a decision to initiate a new drug development project or to initiate GLP safety assessment of a new drug candidate);
- Studies for which any outcome would be considered diagnostic evidence about a claim from prior research (Nosek and Errington, 2020);
- Labor-, resource- and/or time-intensive studies that cannot be easily repeated.

EQIPD requires that investigators assert in advance whether a study will be conducted to inform a formal knowledge claim (core requirement #10), and that they explicitly state this in the study (experimental) plans prepared before studies and experiments are conducted.

Further, it is required for all types of research that everyone in the research unit is adequately trained and competent (core requirement #11), has access to protocols for experimental methods (core requirement #12), follows adequate procedures for the handling and storage of samples and materials (core requirement #13), and uses research equipment and tools that are suitable for the intended use (core requirement #14).

A SYSTEM, NOT JUST A COLLECTION OF GUIDELINES AND RECOMMENDATIONS

Development and implementation of flexible and fit-for-purpose solutions are usually enabled by introducing a continuous improvement process (Deming, 1986). Within the EQIPD environment, the improvement cycle is rooted in the following workflow:

- Understand the rationale for introducing something new or modifying the current work routine (Why - the Need);
- Understand what is needed to achieve it (What - the Challenge);
- Propose a solution for achieving it (How - fit-for-purpose Solution);
- Evaluate the success of the implementation (Assessment).

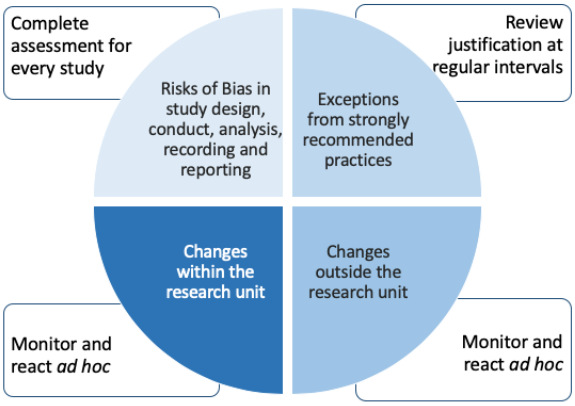
As an example, a research organization is seeking a collaboration with a biopharmaceutical company (Why). The company informs the research organization about its expectations regarding the raw data record generation, handling and storage. The research organization recognizes challenges associated with the storage of raw data as defined by the company (What). The EQIPD Toolbox provides information on what is the raw

data and what are the best practices in recording and handling the raw data (How). In many cases, the new way of working is applied and has the desired effect. In some cases, there may be deficiencies identified that require remediation such as changes in the protocols, additional communication, educational and training efforts. Evaluation of the success in implementation of new processes concludes the cycle (Assessment).

In addition, the successful use of a new method or procedure often requires training, adequate and timely communication, feedback on incidents and errors, etc. To fully establish the EQIPD Quality System, several corrective or feedback mechanisms have to be included. These mechanisms identify factors affecting the generation, processing and reporting of research data *before* a study is done (core requirement #15; see Box 1), to analyze and manage the incidents and errors that may occur *during* the study (core requirement #16), and to monitor the performance of the EQIPD Quality System (core requirement #17; see Box 2).

BOX 1 --- Managing risks to data quality

Even under the best circumstances, not all recommended practices and protection mea-



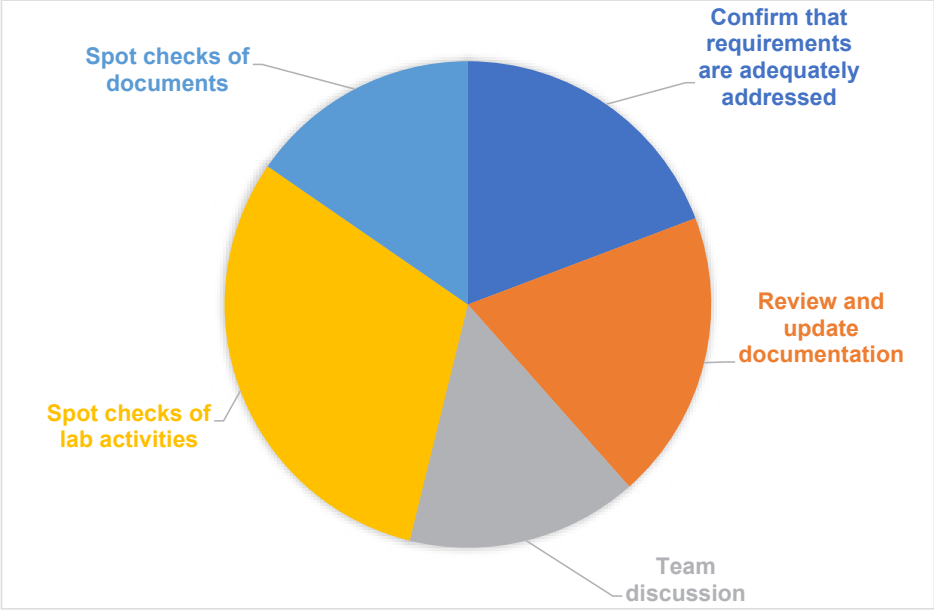
sures can be applied to a working environment or research study, leaving a potential risk of failure. The EQIPD Quality System recognizes the following main areas where risk assessment should be conducted with risks made transparent and, if appropriate, documented:

1. alterations from strongly recommended practices (i.e., situations in which the language of the EQIPD guidance includes "should" and the research unit justifies why it does not or cannot apply). These assessments are done at regular intervals by the Process Owner;

2. key and support processes that are inherently associated with risks endangering the validity of the results (e.g., risk of unblinding; emergency access to blinding codes). These assessments are done by scientists responsible for a study plan;
3. changes in the environment both inside and outside of the research unit (colleagues leaving; facility changes, etc.). These assessments are done or initiated *ad hoc* by the Process Owner.

BOX 2 --- Self-assessment

The primary objectives of the self-assessment are to confirm that the research unit has everything in place for proper performance of the fit-for-purpose EQIPD Quality System,



and to set the basis for internal or external quality checks / accreditation mechanism.

The process owner is responsible for defining the scope and frequency of this self-assessment, which is expected to involve all members of the research unit to ensure that all quality goals in the research unit have been considered and achieved.

As part of the self-assessment, there are spot checks conducted on selected documents (core requirements ## 11, 12, 16, 17; Table 2) and laboratory activities (core requirements ## 6, 7, 8, 9, 10, 13, 14, 15). The Process Owner completes a paperless assessment of several solutions being up-to-date (core requirements ## 1, 2, 4, 5), reviews and, if neces-

sary, updates documentation (core requirements ## 2, 3, 6, 7, 8), and engages the team in the discussion and review of certain processes (core requirements ## 3, 5, 13, 16). The self-assessment itself is a core requirement (#17) and can be conducted using a template provided in the Toolbox.

DEFINING THE USER OF THE EQIPD QUALITY SYSTEM

The ultimate mission of the EQIPD Quality System is to serve the entire community of scientists conducting non-regulated preclinical biomedical research. To achieve this goal, EQIPD has developed and is executing a dissemination strategy that will *initially* focus on early adopters, i.e., research groups and scientists who:

1. See the value of higher standards of rigor in research to achieve more robust and reliable results, are willing to learn about and adopt a quality mindset and are prepared to invest effort to set up the EQIPD Quality System;
2. Consider their standards of rigor are already good, but strive to improve them further, and would like to establish the EQIPD Quality System as an independent *seal of quality*;
3. Can use the EQIPD Quality System to strengthen a grant application, to support decision-making in drug discovery and /or to promote their services (e.g., CROs or academic labs active in the contract research domain) and bolster their reputation;
4. Are motivated by their funders, publishers and collaboration partners to secure high-rigor research standards (e.g., as a condition for funding or collaboration).

Such early adopters are known to be of critical value in every field where a cultural change is under discussion. For instance, academic initiatives have successfully addressed research data management and sharing of best practices by introducing Data Champions that serve as local advocates for good data practices (e.g., https://www.data.cam.ac.uk/intro-data-champions). Peer-to-peer learning eventually supports the dissemination of good practices beyond the early adopters.

The early adopters of the EQIPD Quality System, through their feedback to the EQIPD consortium, will help optimize the balance between the benefits of implementing such a system and any potential adverse consequences (e.g., resources allocated, reduction in conventional indices of scientific productivity). A positive balance will support further dissemination of the EQIPD Quality System and help broader research communities take advantage of the work done by the EQIPD team and the early adopters.

It is a general understanding that not all research units are equally prepared or will be willing to implement a Quality System, an effort that requires investing time and institutional resources. EQIPD has developed and shared resources relating to the quality system that can be used for other purposes – as a source of information about specific aspects of good research practice, as a guidance for specific types of projects (e.g., industry-academia collaboration), or to enable a specific collaboration project by providing a purpose-fit certification of the current practices being in line with the EQIPD expectations (Table 5).

Since the scientists themselves will be the main users of the EQIPD framework, their leading and proactive role in improving the quality of their own scientific data will define

TABLE 5 --- Levels of use of the EQIPD framework

Levels of use:	Information only (incl. training)	Purpose-fit certification	Quality System
EQIPD guidance:	Recommendations on best practices, examples, templates	Basic set of core requirements	Full set of core requirements
Main users:	Research units, funding organizations	Research units	Research units
Expected use:	As necessary, follow specific recommendations or use provided tools to improve work processes (e.g., increase transparency or make raw data findable or improve reporting) As appropriate, use information provided by EQIPD in training programs; communicate to collaborators, grantees, etc.	Confirm that current quality practices are in line with the basic set of EQIPD core requirements (related to data integrity and rigor in study design, conduct, analysis, and reporting)	Align current research quality practices with the EQIPD expectations (implement full set of core requirements including those that define quality system – i.e., availability of resources, process owner, quality objectives, and continuous improvement mechanisms)
Dedicated efforts by the research unit (e.g., regular and sustained efforts, dedicated personnel)	None	Limited	Yes
Context of use	Research unit is informed about expectations by current or future collaborators, funders, sponsors, publishers, etc.	Flexible solution driven by the time- and resource-critical needs of specific collaboration(s)	Stable solution for long-term maintenance of research rigor standards
Assessment by the EQIPD team	No	Yes	Yes

the ways the framework can be used to prepare more and more research units to accept a Quality System as a means for long-term maintenance or research rigor standards.

IMPLEMENTATION OF THE EQIPD QUALITY SYSTEM

Even a lean and user-friendly quality system requires effort and resources to be implemented and maintained. This consideration makes it important to emphasize that a decision to start implementing the EQIPD Quality System should be well justified and regularly checked by the Process Owner and discussed with the research team.

Size of the research unit

Ideally, the EQIPD Quality System should be implemented at the level of an organization (university, research institute, or a company). While this is the desired case, EQIPD encourages the transition towards better quality practices at the level of individual labs, departments or research groups, no matter how small they are, provided that there is a researcher capable, authorized and willing to take on the role of Process Owner.

The EQIPD Quality System is not intended to be used at the level of individual projects. Otherwise, it may create confusion and increase the risk of errors as the same people within a research unit may follow separate research quality practices depending on the project that they are working on.

Implementation path

There are several ways in which the EQIPD core requirements can be introduced within a research unit in terms of timing and sequence (Figure 2). Whether supported by the (optional) EQIPD tools or not, any of the possible implementation scenarios are acceptable as long as the outcome is the same – that is, a quality system implementing all 18 core requirements.

FIGURE 2

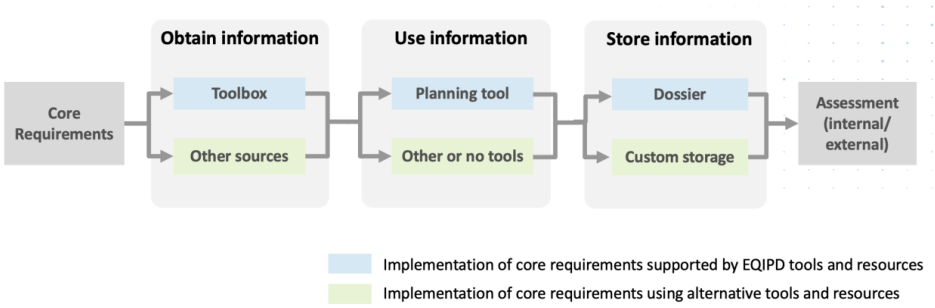


Figure 2: Implementation of the EQIPD Quality System (QS): From Core Requirements (CR) to assessment of a fully functional system.

The 18 CRs are the expectations formulated by the EQIPD that serve as the starting point for implementing the QS. At any step during the implementation, the use of EQIPD tools is voluntary and serves only the purpose of making the implementation and maintenance of the QS easier. As the first step, unless such information is available from other sources, the research unit may consult with the Toolbox to obtain relevant research quality-related information. Once the necessary information is obtained, the research unit applies this knowledge and monitors the progress. This can be done using the Planning Tool, using alternative project management resources or even without any such tools. The Dossier is a repository of documents and information that are specific to the user's research unit and that is organized according to a structure suggested by EQIPD (to keep all research quality-related information in one place and make it easily findable). However, the research unit may also opt to use its own way to store information. Finally, once the implementation is completed, the research unit may initiate an assessment to get feedback from experts outside of the research unit (either quality professionals within the same organization or a third party).

The implementation path suggested by EQIPD envisions three phases (Supplement S3 Implementation path):

Phase 1– A short list of cornerstone actions that are the same for all research units to help users understand why things are done, as well as ensuring that efforts triggered by the EQIPD framework have immediate impact (e.g., best practices to support data integrity and traceability).

Phase 2 – Users develop solutions for challenges directly connected to their environment or needs communicated by their funders, publishers and collaboration partners. During this phase, users meet most of the EQIPD core requirements while developing a habit of working towards a quality system.

Phase 3 – Completion of the remaining core requirements enabling formal recognition of a functional quality system.

The implementation is concluded with an important sustainability checkpoint: the Process Owner is expected to estimate the required resources and make them available for maintaining the EQIPD Quality System (core requirement #18).

Supporting tools

EQIPD has developed several tools (Figure 2) to support the implementation and maintenance of the Quality System:

- The Toolbox is a structured collection of information that enables users to build or select solutions for customized research needs. This Toolbox is built using wiki principles. The Toolbox contains a growing body of information about existing guidelines, recommendations, examples, templates, links to other resources, literature references, or just guidance on how to address a specific topic and will be regularly updated.
- The Planning Tool is a user interface, designed to review the needs of researchers and is specific to their environment and focus of their research. Summarized expectations of funders, publishers, and collaboration partners can be entered in the Planning Tool either directly or using a special template called the Creator Tool.
- The Dossier is a structured collection of customized documents and information related to research quality in a given research unit.

EQIPD does not intend to insist that researchers use these tools and rather sees their application as optional.

THE EQIPD QUALITY SYSTEM: COMPLIANCE MECHANISMS

The EQIPD system is a voluntary quality assurance framework that enables research units to demonstate compliance with 18 core requirements which can fulfill their own quality needs, e.g., community guidelines or funder requirements.

Traditional quality systems require either internal (within the organization) or external auditors to check compliance with its system. This in turn requires that organizations employ dedicated and adequately trained quality professionals that understand the specific language in these quality regulations and ensure that the documentation formats correspond to the norm and nomenclature of the certifying organization.

EQIPD Quality System is conceived as beginning with research scientists and extending to the research environment, and the compliance mechanisms are in line with this approach.

Self-Assessment

The process owner is expected to use a self-assessment form provided by EQIPD to check whether Core Requirements and research unit-specific needs are appropriately addressed. The form guides the process owner through each core requirement, links out to the corresponding online Toolbox item, which describes background, expectations and provides further guidance documents.

The self-assessment serves two purposes. On the one hand, it allows the process owner to monitor performance of the quality system. On the other hand, it provides the base for an external assessment.

External assessment

External assessors review the self-assessment document and may request the research unit to provide additional documentation, such as training materials, research output, relevant parts from a study plan, raw data entries, or animal care and use procedures.

Assessors review the documents and, based on the information provided, decide whether each core requirement is sufficiently addressed or whether additional verification is needed during the assessment interview. To aid the decision-making process, a reviewer guidance document contains specific questions for each core requirements that may be already addressed in the material provided or need further verification.

The results and questions of this pre-assessment are shared with the research unit and are discussed in detail and clarified during the subsequent interview. These interviews can be performed at the research site or via online videoconferencing. A report is prepared by the assessors that details the results of the assessment, contains suggestions for improvement and ultimately confirms whether the research unit is compliant with all core requirements or not yet. The assessed research unit is given the opportunity to respond and to suggest changes to the report before it is finalized.

Research units that successfully implemented the EQIPD Quality System receive a certificate of EQIPD compliance.

Several research units have completed the implementation of the EQIPD Quality System and have already been evaluated by the EQIPD team.

External assessment is currently performed by scientists that developed the EQIPD Quality System. A training modules for future assessors will be released to ensure the reliability and consistency of assessments conducted by different experts.

Moreover, anticipating a large demand for external assessments, the EQIPD team evaluates and compares the reliability of various external assessment models combining onsite visits and remote interviews.

Importantly, EQIPD aims to make the assessment process as straightforward as possible. EQIPD's expectations are concisely summarized for each core requirement in a document that is regularly updated and available via the Toolbox. Further, the EQIPD team advises to refer to the five key principles (Table 3) whenever a specific answer is not yet provided in the EQIPD guidance.

Last but not least, EQIPD's vision is that the Quality System serves the research units in the role of a partner, stimulating and guiding the continuous improvement in research rigor. With that in mind, EQIPD places a lot of weight on the competence and engagement of process owners conducting regular spot checks of key research processes and documentation.

ENHANCING QUALITY IN PRECLINICAL DATA (EQIPD): THE OUTLOOK

On September 30, 2020, the EQIPD Quality System was released for broad deployment and unrestricted use by the research community.

To enable the maintenance and further development of the EQIPD framework beyond the IMI project phase, the EQIPD team is implementing a governance model (Figure 3). The proposed model comprises three closely interacting levels:

- A strategic level represented by the EQIPD Guarantors, a group of the EQIPD project team members responsible for the overall guidance, administration of academic and educational programs, and the dissemination of the EQIPD vision. The EQIPD Guarantors will be supported by an Ethics & Advisory Board, a consultative body composed of current EQIPD consortium members, associate collaborators and advisors as well as key opinion leaders in the field of good research practice.
- An operational level represented by an independent globally acting partner organization, commissioned by the EQIPD Guarantors to provide the operational support and services required for day-to-day business management (including technical support and training for the research units during the implementation and maintenance of the EQIPD Quality System).
- A community level that is represented by the EQIPD Stakeholder group, a diverse group of scientists, funders, quality professionals, manufacturers of research tools,

and publishers that provide feedback on practical aspects of the EQIPD Quality System and facilitates connections to a broader biomedical research community.

FIGURE 3

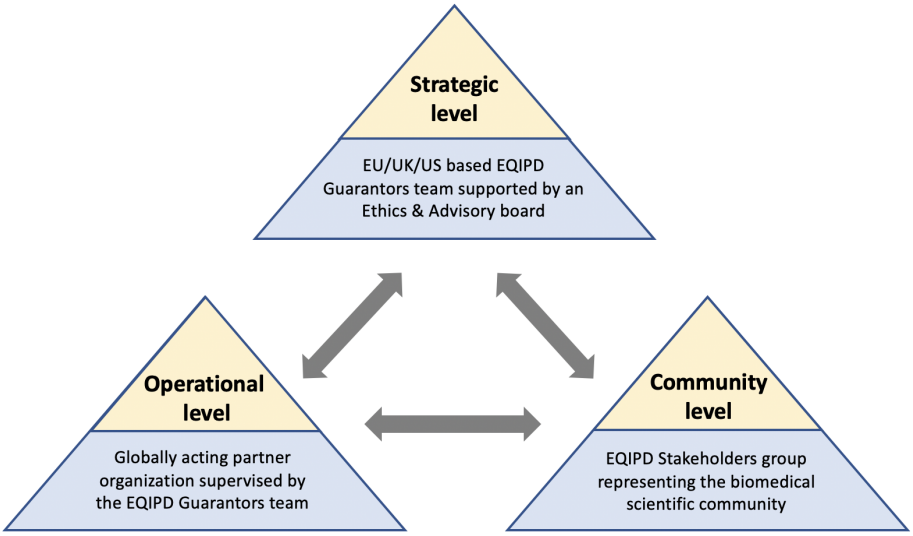


Figure 3: The proposed future governance model of EQIPD. The EQIPD Guarantors group and the EQIPD Ethics & Advisory Board are responsible for the overall guidance, administration of academic and educational programs, as well as dissemination of the EQIPD vision (*Strategic level*). An independent partner organization, commissioned by the EQIPD Guarantors, will provide the operational support and the day-to-day services for the EQIPD community (*Operational level*). The EQIPD Stakeholder group, composed of scientists, funders, quality professionals, manufacturers of research tools, and publishers, provides feedback on the practical aspects of the EQIPD Quality System and facilitates connections to a broader biomedical research community (*Community level*).

The next milestones for the EQIPD team are:

- Launch of an educational platform that will support both the use of the EQIPD Quality System and provide more general training in the field of good research practice;
- Analysis of geographical and cultural differences that may affect the acceptance of the EQIPD Quality System and that may require adaptations in the associated framework;
- Evaluation of the impact of implementation of the EQIPD Quality System on research quality, to inform further development of the EQIPD framework.

The EQIPD Quality System was developed with the focus on the users and their needs. The EQIPD collaborators will maintain and expand this focus further.

The EQIPD team is actively engaged in discussions with funders (public and private) and publishers to develop instruments and mechanisms that will allow scientists to further benefit from the use of the EQIPD Quality System.

All scientists engaged in preclinical biomedical research are invited to join the growing community of the EQIPD Quality System users and supporters (www.eqipd.online).

ACKNOWLEDGMENTS

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777364. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

The authors are very grateful to Martin Heinrich (Abbvie, Ludwigshafen, Germany) for the exceptional IT support and programming the EQIPD Planning Tool and the Creator Tool and to Dr Shai Silberberg (NINDS, USA), Dr. Renza Roncarati (PAASP Italy) and Dr Judith Homberg (Radboud University, Nijmegen) for highly stimulating contributions to the discussions and comments on earlier versions of this manuscript. We also wish to express our thanks to Dr. Sara Stöber (concentris research management GmbH, Fürstenfeldbruck, Germany) for excellent and continuous support of this project.

Creation of the EQIPD Stakeholder group was supported by Noldus Information Technology bv (Wageningen, the Netherlands).

DISCLOSURES

AB, RB, AG, BG, VC, IAL, FD, LM, SB, BA, MAA, CE, CFC, EH, MJK, CK, MP, HP, GR, MRH, JV, KW, MM, UD, and TS are current or past employees of the organizations that are founding members of the EQIPD consortium. JG is an employee of AAALAC International that is an EQIPD Associated Collaborator. LB, NdB, AD, RF, CFB, VGD, SMH, MH, PK, PM, PP, EPMdO, LR, JS, MS, CS, and VV are members or are current or past employees of the organizations that are members of the EQIPD Stakeholder group. AB is an employee and/or shareholder at PAASP GmbH, PAASP US LLC, Exciva GmbH, Synventa LLC, Ritec Pharma. AB, BA, NdB, UD, CFB, PK, MK, MM, PM, PP, GR, JS, and TS are members of the Preclinical Data Forum (co-chairs – AB and TS), a network financially and organizationally supported by ECNP and Cohen Veterans Bioscience. LM is an employee and shareholder of UCB. SB is an employee of Novartis Pharma. HP has received during the last three years consulting

and speaking fees and/or funding for collaborative projects from Bayer, Roche, Zogenix, and Eisai. IAL and FD are employees of Sanofi. BA is an employee and shareholder of Pfizer. The views and opinions expressed in this article are those of the individual author and should not be attributed to Pfizer, its directors, officers, employees, affiliates, or any organization with which the author is employed or affiliated. VC and CFC are employees of Porsolt. PK is an employee and shareholder at PAASP US LLC. PM is owner of Cerbascience Consulting. UD and CK receive funding from Volkswagen Foundation. BG and CE are employees and shareholders at PAASP GmbH. MM, UD and TS are members of the Advisory Board at PAASP. MM, UD and TS are members of the ARRIVE guidelines working group. KW is a consultant of Avertim, Brussels, Belgium, support for this contribution was funded by Janssen Pharmaceutica NV. TS is an AAALAC ad-hoc specialist. TS and AG are employees of Janssen / Johnson & Johnson and shareholders at Johnson & Johnson.

REFERENCES

1. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circ Res. 2015; 116(1): 116-26. doi: 10.1161/CIRCRESAHA.114.303819.
2. Bespalov A, Steckler T, Altevogt B, Koustova E, Skolnick P, Deaver D, Millan MJ, Bastlund JF, Doller D, Witkin J, Moser P, O'Donnell P, Ebert U, Geyer MA, Prinssen E, Ballard T, Macleod M. Failed trials for central nervous system disorders do not necessarily invalidate preclinical models and drug targets. Nat Rev Drug Discov. 2016; 15(7): 516. doi: 10.1038/nrd.2016.88.
3. Bongiovanni S, Purdue R, Kornienko O, Bernard R. Quality in non-GxP research environment. Handb Exp Pharmacol. 2020; 257: 1-17. doi: 10.1007/164_2019_274.
4. Deming WE (1986). Out of the crisis. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study. 524 p.
5. DiMasi JA, Grabowski HG, Hansen RW (2016) Innovation in the pharmaceutical industry: New estimates of R&D costs. Health Econ. 2016; 47: 20-33. doi: 10.1016/j.jhealeco.2016.01.012.
6. Dirnagl U. Thomas Willis lecture: Is translational stroke research broken, and if so, how can we fix it? Stroke. 2016; 47(8): 2148-2153. doi: 0.1161/STROKEAHA.116.013244.
7. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol. 2015; 13(6):e1002165. doi: 10.1371/journal.pbio.1002165.
8. Gilis A. Quality governance in biomedical research. Handb Exp Pharmacol. 2020; 257: 349-365. doi: 10.1007/164_2019_291.
9. Guillén J. The use of performance standards by AAALAC International to evaluate ethical review in European institutions. Lab Anim. 2010; 39: 49-53. https://doi.org/10.1038/laban0210-49.
10. Guillén J, Steckler T. Good Research Practice: Lessons from animal care and use. Handb Exp Pharmacol. 2020; 257:367-382. doi: 10.1007/164_2019_292.
11. Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. BMC Med Res Methodol. 2014; 14: 43. doi: 10.1186/1471-2288-14-43.
12. Juran JM, Godfrey AB. Juran's Quality Handbook, 5th Edition, 1999, McGraw Hill Professional, 1999, 1872 p.
13. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. PLoS Biol. 2014; 12(5): e1001863. doi: 10.1371/journal.pbio.1001863.
14. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, Finkelstein R, Fisher M, Gendelman HE, Golub RM, Goudreau JL, Gross RA, Gubitz AK, Hesterlee SE, Howells DW, Huguenard J, Kelner K, Koroshetz W, Krainc D, Lazic SE, Levine MS, Macleod MR, McCall JM, Moxley RT 3rd, Narasimhan K, Noble LJ, Perrin S, Porter JD, Steward O, Unger E, Utz U, Silberberg SD. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012; 490(7419): 187-91. doi: 10.1038/nature11556.
15. Marshall E. The murky world of toxicity testing. Science. 1983; 220(4602): 1130-1132.
16. Nosek BA, Errington TM. What is replication? PLoS Biol. 2020; 18(3): e3000691. doi: 10.1371/journal.pbio.3000691.
17. Ritskes-Hoitinga M, Wever K. Improving the conduct, reporting and appraisal of animal research. BMJ. 2018; 360: j4935. doi: 10.1136/bmj.j4935.
18. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. 2012; 11(3): 191-200. doi: 10.1038/nrd3681.

19. Shih HP, Zhang X, Aronov AM. Drug discovery effectiveness from the standpoint of therapeutic mechanisms and indications. Nat Rev Drug Discov. 2018; 17(1): 19-33. doi: 10.1038/nrd.2017.194.
20. Vollert J, Schenker E, Macleod M, Bespalov A, Wuerbel H, Michel M, et al. Systematic review of guidelines for internal validity in the design, conduct and analysis of preclinical biomedical experiments involving laboratory animals. BMJ Open Science 2020; 4: e100046. doi: 10.1136/bmjos-2019-100046.
21. Volsen SG, Masson MM. A novel audit model for assessing quality in non-regulated research. Qual Assur J 2009; 12: 57–63. doi: 10.1002/qaj.441.
Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM, Pairaudeau G, Pennie WD, Pickett SD, Wang J, Wallace O, Weir A. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat Rev Drug Discov 2015; 14(7): 475

A2

Appendix 2

CONTEXT

From November 2017 to July 2020, we conducted a behavioral and molecular evaluation of the phenotype of the rTg4510 tau-pathology model. This model has been widely used as a proxy for Alzheimer's disease given that the mouse line expresses a form of human tau (tau P301L) ("MAPT P301L | ALZFORUM," November, 2017), in such a way that mice exhibit some of the AD main traits such as memory deficits along with aberrant LTP, and mislocalization of tau to somatodendritic space and dendritic spines, presumably responsible for synaptic dysfunction (Hoover et al., 2010). Therefore, we wanted to explore the progress of the pathology from early stages in the development to when the pathology is already established, typically characterized at this stage. The aim was to explore the presence of early signs of cognitive decline as a possible early indicator that the pathology was developing; this with the aim to find an early intervention. However, close to the final step of data recollection a publication came out showing how the transgenic mutation had consequences in other genes. The implications of this made it impossible to disentangle whether phenotype expressed was due to the tau-pathology or to any other of the mutations since these also had functional roles in the brain. The project stopped there, however most of the behavioral data was already collected and analyzed. Beyond the impossibility to interpret the results in light of translational validity, results showed conflicting alignment with literature in terms of the consistency of the behavioral phenotype. Although the study was not meant to be a direct replication of previous published studies, it highlights how independent studies using the same study population with similar methodologies can have inconsistent results.

Figure 1 shows the behavioural phenotype results of the Tg4510 mouse model for previously publishes studies (Foster et al., 2019; left column) compared to in-house results (right column). **Panel A** shows short-term memory performance in the Y-maze (left) found a short-term memory deficit early on (4 weeks) that persisted across development (8 and 12 weeks). On the contrary, we did not find a memory deficit at 14 weeks; strikingly, we found an enhanced performance on behalf of the Tg4510 mice at 22 weeks of age. Our finding was attributed to the enhanced locomotion showed by the Tg4510 (see panel B), however, we took the same criteria as Foster and colleagues to exclude those animals that showed >90% alternations. **Panel B** shows the locomotor activity in the open field; both sites consistently found a hyperactive phenotype for the Tg4510 mice from 14 to 30 weeks of age. **Panel C** shows the performance in the Barnes maze for Foster and colleagues, and modified Barnes maze (#) for Tg4510 mice and their controls. Foster reported statistical difference for the time spent in the target quadrant, meaning the Tg4510 mice showed a memory deficit. On the other hand, our results showed no deficit for the Tg4510 mice. In addition, the accuracy of performance of both experimen-

tal groups in our site is higher than in Foster's with animals spending ~50% of the time in the target quadrant. This could suggest an overtraining of the experimental animals at our side, although we followed the same protocol as Foster with a more difficult setup. The modified Barnes Maze, in comparison to the regular Barnes Maze used by Foster, has 42 holes instead of 17 scattered around the tables, instead of only around the edge. This makes the task more difficult, thus, the same amount of training should not lead to overtrained animals (ceiling effect).

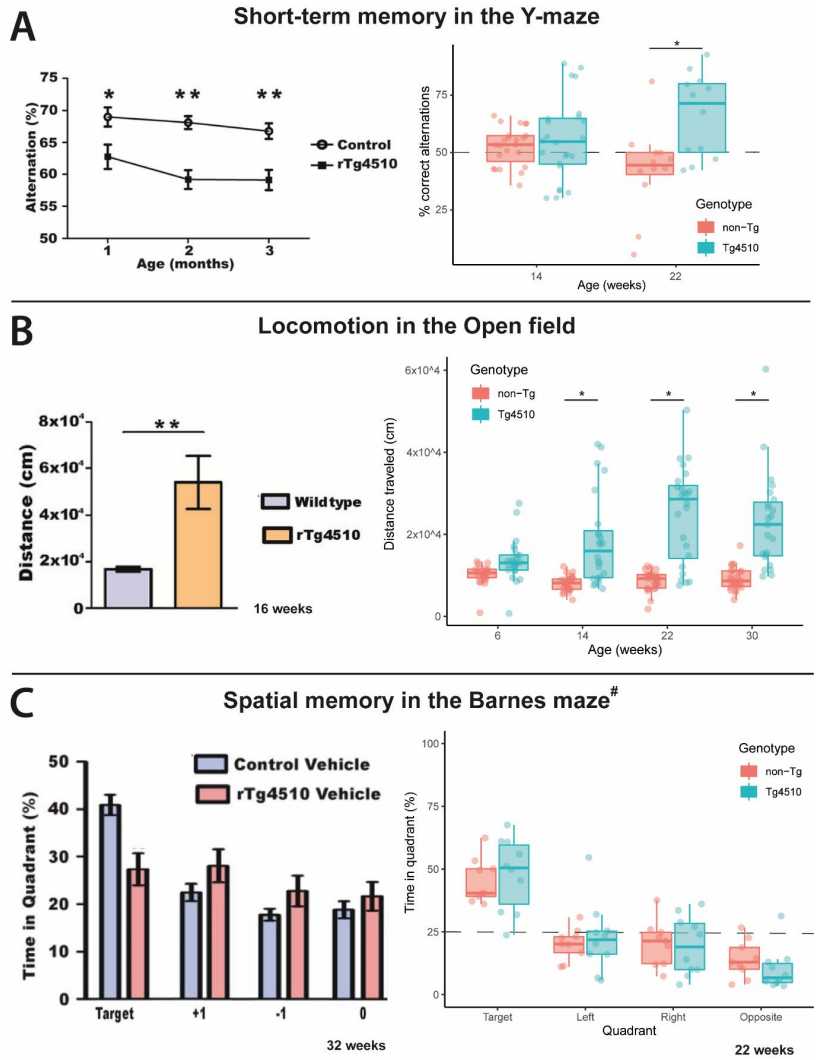


Figure 1. Comparison of behavioral results from Foster et al., (2019) and our lab (right panel) for the short-term memory evaluation (Panel A), locomotion in the open field (Panel B) and Spatial memory in the Barnes maze for Foster and modified Barnes maze for our lab (Panel C).

Overall, these results highlight the variability of results depending on the robustness of the phenotype and likely on the complexity of the task. In this case, the open field test is the simplest and apparently the most robust since albeit the differences in methodologies, both sites found hyperlocomotion. However, for the Y-maze alternations and the Barnes maze, there may be methodological differences that can be causing the discrepancies although this cannot be assessed directly by reading the published results from Foster and our methods.

METHODS

Animals

rTg4510 and non-transgenic littermates (non-Tg) were bred as described previously (SantaCruz et al., 2005). All animals were group housed by sex in different rooms with three or four littermates. The cages had shelter, nesting material and a paper roll for enrichment. Food and water were available *ad libitum*. A 12hr light-dark cycle was followed with lights turning on at 01:00. Constant temperature (~21°C) and relative humidity (~50%) was observed in the nest room as well as across all experiments. Animal cages were cleaned once a week and had their weight monitored through the experiments. We tested mice from different ages (6, 14, 22 and 30 weeks old), coming from different cohorts.

Behavioural testing

Experimenters were blinded to the genotype of the animals throughout the experiment and analysis. Animals ere tested in a randomized order. During experiments, male and female animals were tested in separate rooms. G-power analysis indicated that a sample size of 24 per genotype, per timepoint would be the minimum size to detect an effect size. This was balanced for sex with 12 female and 12 male animals being tested at each timepoint. Loss of animals occurred during the testing due to the sacrificing of males as they tended to fight. Animals for each timepoint were independent batches that ran the tests. The spontaneous alternations in the Y-maze were not tested in 6- and 30-weeks old mice given the high rate of stereotypic behaviour shown in this test. Instead, the Novel object location was used, however the protocol was sub-optimal as not even the control littermates had a performance above chance levels, thus data is not shown. For the Barnes Maze, only one time-point is exhibited here, however, the other timepoints did not show a phenotype effect. The tests were always conducted during the dark phase of the light-dark cycle and in the order listed below.

Spontaneous alternations

The Y-maze was used to assess working memory by means of spontaneous alternation in a Y-maze. For this, animals were habituated to the test room for 30mins before undergoing the trial under red light conditions. A mouse was placed in the centre of the Y-maze with a lid on top of the maze to prevent the mouse to jump out. The video recording was run for 10 minutes, after which the mouse was removed and the arena cleaned with 0.1% acetic acid. The arms of the maze were labelled A, B, C. An entry was recoded when at least 3 paws of the mouse were inside the arm. Number of correct alternations was automatically scored using EthoVision (Noldus Information Technologies, Wageningen, NL) which was defined as 3 consecutive arm entries to different arms without repetition ((Blackmore et al., 2017). This value was then divided by the total number of alterations and multiplied by 100 to get a percentage of correct alternation which was the main outcome.

Open Field Test

The open field test was used to test for spontaneous locomotion in the animals. The arena used was 80cm in diameter with walls 30cm tall. A white PVC bottom was used to better visualise the animals. Before beginning the test, animals were acclimatised to the testing room in single housing for 30 minutes before the start of the experiment. At the start of the test, the animal was placed in the centre of the arena and video recorded for 15mins before being removed and the arena cleaned with 0.1% acetic acid. The data was analysed for distance travelled and velocity using the EthoVision software (Noldus Information Technologies, Wageningen, NL).

Modified Barnes Maze

The Barnes maze was used to test spatial memory in the animals. This arena is modified from the original version to reduce the likelihood of the animal finding the box by chance by the addition of more holes that are scattered across the whole arena instead of sequentially around the edge of the arena. Animals were trained for 7 days, two trial per day. Each trial consisted of placing the animal inside a cylinder on the centre of the arena, the experimenter pulled a chord to lift the cylinder and release the mouse from outside the experimental room. The mouse had to navigate the maze according to cues located equidistant from the maze to find the hole that contained the escape box. Animals had up to 5 minutes to find the escape box, otherwise it was gently guided to the escape box. Once inside the escape box, it was transferred back to its' cage. All animals were previously habituated to a box identical to the escape box in their home cages at least 48 hours in advance to the start of the training. On the last session (probe trial), the escape box was removed and the animal was left to explore the maze for 5 minutes. The main outcome was the percentage of time spent in the quadrant where

the escape box was (target quadrant) compared to the other quadrants of the table. This was obtained from the probe trial.

Statistical analysis

Y-maze alternations

As reported by Foster, we excluded those animals with alternation percentage >90%. We only found one female Tg4510 animal of 22 weeks of age with 92% which left the total number of animals in 70 (35 Tg4510, 35 non-Tg, 50% females; 47 were 14 weeks old, 23 were 22 weeks old).

The following general linear model was fitted to the data: alternation % ~ Genotype * Age. This model reported a significant interaction where there was genotype effect found only in the 22 weeks old mice (Table 1). Tg4510 animals of 22 weeks showed increased % of correct alternations compared to their littermates. There was an exploratory analysis on sex but it yielded no difference between sexes nor an improvement in the model (based in the r squared), therefore this factor was taken out of the final model for parsimonious purposes.

Table1. Statistical summary of the interaction between Genotype and Age for spontaneous alternations in the Y-maze

	Estimate	Std. Error	T value	Pr(>|t|)
Intercept	52.803	3.067	17.219	< 2e-16 ***
Genotype: Tg4510	2.855	4.291	0.665	0.50817
Age: 14 weeks	-13.785	5.391	-2.557	0.01290 *
Genotype X Age	22.123	7.599	2.911	0.00492 **

Residual standard error: 14.71 on 65 degrees of freedom; Adjusted R-squared: 0.1691

Open field test

The distance travelled data was converted to natural logarithm to comply with normality and homoscedasticity criteria. There were 190 animals in total; 46 at 6 weeks of age, 48 at 14 weeks, 49 at 22 weeks and 47 at 30 weeks; 50% females). The following model was fitted to the data: log (Distance travelled) ~ Genotype*Age. There was an exploratory analysis on sex but it yielded no difference between sexes nor an improvement in the model (based in the r squared), therefore this factor was taken out of the final model for parsimonious purposes. The final analysis showed in Table 2 had a significant interaction of Genotype X Age where the genotype effect across ages was different as visible in Figure 1.

Table 2. Statistical summary on the interaction between Genotype and Age for the distance travelled in the OF.

	Estimate	Std. Error	T value	Pr (>|t|)
Intercept	9.1469	0.1047	87.329	< 2e-16 ***
Genotype: Tg4510	0.2504	0.1450	1.727	0.085873
Age: 14 weeks	-0.1859	0.1450	-1.282	0.201497
Age: 22 weeks	-0.1426	0.1436	-0.993	0.322065
Age: 30 weeks	-0.0676	0.1450	-0.466	0.641653
Genotype X 14 wks	0.4559	0.2028	2.248	0.025799 *
Genotype X 22 wks	0.7335	0.2018	3.634	0.000363 ***
Genotype X 30 wks	0.6040	0.2039	2.962	0.003461 **

Residual standard error: 0.4913 on 182 degrees of freedom; Adjusted R-squared: 0.3744

Modified Barnes Maze

The natural logarithm of the time spent in the different quadrants was fitted to the following model: log (Time in quadrant) ~ Genotype + Quadrant. There were 80 female mice with 22 weeks old; 40 Tg4510 and 40 non-Tg. Table 3 shows there was an overall significant difference of the target quadrant compared to the other 3 quadrants. Meaning that Tg4510 and non-Tg animals learned the spatial location of the escape box accurately.

Table 3. Statistical summary of the Genotype effect on the target quadrant time in the modified Barnes Maze

	Estimate	Std. Error	T value	P(>|t|)
Intercept	3.8895	0.1424	27.319	< 2e-16 ***
Genotype: Tg4510	-0.1474	0.1261	-1.168	0.247
Quadrant: Left	-0.8728	0.1781	-4.900	5.84e-06 ***
Quadrant: Opposite	-1.5445	0.1781	-8.671	9.53e-13 ***
Quadrant: Right	-0.9807	0.1781	-5.506	5.53e-07 ***

Residual standard error: 0.549 on 71 degrees of freedom; Adjusted R-squared: 0.4981

REFERENCES

Blackmore, T., Meftah, S., Murray, T. K., Craig, P. J., Blockeel, A., Phillips, K., Eastwood, B., O'Neill, M. J., Marston, H., Ahmed, Z., Gilmour, G., & Gastambide, F. (2017). Tracking progressive pathological and functional decline in the rTg4510 mouse model of tauopathy. Alzheimer's Research and Therapy, 9(1), 1–15. https://doi.org/10.1186/s13195-017-0306-2

Foster, J. B., Lashley, R., Zhao, F., Wang, X., Kung, N., Askwith, C. C., Lin, L., Shultis, M. W., Hodgetts, K. J., & Lin, C.-L. G. (2019). Enhancement of tripartite synapses as a potential therapeutic strategy for Alzheimer's disease: A preclinical study in rTg4510 mice. Alzheimer's Research & Therapy, 11(1), 75. https://doi.org/10.1186/s13195-019-0530-z

SantaCruz, K., Lewis, J., Spires, T., Paulson, J., Kotilinek, L., Ingelsson, M., Guimaraes, A., DeTure, M., Ramsden, M., McGowan, E., Forster, C., Yue, M., Orne, J., Janus, C., Mariash, A., Kuskowski, M., Hyman, B., Hutton, M., & Ashe, K. H. (2005). Tau Suppression in a Neurodegenerative Mouse Model Improves Memory Function. Science (New York, N.Y.), 309(5733), 476–481. https://doi.org/10.1126/science.1113694

Nederlandse samenvatting

Het kunnen verkrijgen van vergelijkbare resultaten bij herhaling van een experiment met vergelijkbare methodiek wordt gezien als een manier om de waarheidsgetrouwheid van wetenschappelijke bevindingen te bevestigen. Echter, in het afgelopen decennium hebben vele wetenschappelijke publicaties de lage reproduceerbaarheid van resultaten in preklinische studies aangetoond. In reactie hierop zijn verschillende oorzaken van deze zogenoemde replicatiecrisis onderzocht. Hoewel erg divers, zijn de meeste van deze oorzaken geassocieerd met hoe het onderzoek is gepland, uitgevoerd en gerapporteerd; maar de suboptimale onderzoekspraktijken die leiden tot niet-reproduceerbare resultaten gaan verder dan het proces zelf en beslaan ook de stimulans en het beloningssysteem dat wetenschappelijk onderzoek steunt.

Hoewel een deel van de oorzaken en drijfveren zijn geïdentificeerd, is het noodzakelijk om te onderzoeken of verschillende interventies in het onderzoeksproces kunnen leiden tot meer reproduceerbare resultaten. Daarom was het doel van dit proefschrift het verkennen van verschillende methoden die de manier waarop onderzoek wordt uitgevoerd verbeteren met als uiteindelijke doel om datakwaliteit en de reproduceerbaarheid van resultaten te verhogen.

Een methode om de reproduceerbaarheid van resultaten te testen is doormiddel van multicenter studies zoals beschreven in **Hoofdstuk 2 en 3.** De eerste onderzocht hoe consistent het gedragsfenotype van het Shank2 ratmodel voor autisme was. De drie betrokken laboratoria hadden een bijna volledig gelijk protocol, waarin exact dezelfde opstellingen en software werden gebruikt. De resultaten van deze studie lieten zien dat het afstemmen van protocol en opstelling toereikend is om vergelijkbare resultaten te genereren in meerdere laboratoria; echter, de rigoureuze standaardisatie van het protocol limiteert mogelijk de generaliseerbaarheid van de resultaten. Daarentegen, onderzocht H**oofdstuk 3** de reproduceerbaarheid van gedragsresultaten na farmacologische interventies tussen zeven laboratoria met het gebruik van verschillende experimentele designs, relevant voor de generaliseerbaarheid van de bevindingen. Het voornaamste verschil tussen de verschillende experimentele designs was de mate van afstemming in het gestandaardiseerde protocol in en tussen de laboratoria. In het kort, deze studie toonde aan dat het afstemmen van een gestandaardiseerd protocol resulteerde in beter reproduceerbare resultaten tussen laboratoria dan wanneer er minimale standaardisatie was. De resultaten toonden ook aan dat, ondanks standaardisatie van het protocol, subtiele verschillen in experimentele en omgevingsfactoren in elk lab nog steeds variatie tussen de laboratoria introduceerde. Deze variatie tussen laboratoria is mogelijk het gevolg van inherente verschillen tussen de laboratoria die werden versterkt door de strikte standaardisatie. Bovendien kon deze variatie niet worden verklaard door de introductie van systematische variatie in de lichtintensiteit tijdens de test en de tijd

van testen. Derhalve zijn verdere studies nodig om te bepalen welke experimentele en omgevingsfactoren de variatie tussen laboratoria kunnen verhogen om de generaliseerbaarheid van resultaten te verhogen.

Een andere methode om zicht te krijgen op de reproduceerbaarheid van resultaten is het uitvoeren van een systematische review en meta-analyse voor specifieke interventies of fenotypes. Hierin kunnen bevindingen uit de literatuur met elkaar worden vergeleken om het effect van bepaalde variabelen te onderzoeken; het effect van alle geïncludeerde studies wordt samengevat in een algehele effect size. Deze methode is gepresenteerd in **Hoofdstuk 4**, waarin het gedragsmatige fenotype van een genetisch muismodel voor Fragiel-X syndroom (Fmr-1 KO) werd onderzocht. De meta-analyse liet zien dat er voor de meeste van de onderzochte gedragscategorieën grote variatie was in de uitkomsten tussen studies (i.e., lage reproduceerbaarheid van fenotypes tussen studies). De resultaten van de verschillende gedragsfenotypes lieten een mismatch zien tussen de symptomen in het muismodel en patiënten zien; dit suggereert dat de limitaties van het model moeten opnieuw moeten worden geëvalueerd. Daarnaast gebruikten we een methode, aangepast vanuit klinische studies, om het risico op bias te schatten, welke de rapportage van onderzoekspraktijken die de introductie van bias in studies minimaliseren beoordeeld (e.g., blindering, randomisatie, complete rapportage van uitkomstmaten). Deze analyse toonde aan dat een groot deel van de studies het gebruik van methoden om bias te minimaliseren suboptimaal rapporteerde, de interpretatie van resultaten verkregen in dit model moet dus voorzichtig overwogen worden. Door de onderrapportage van omgevingscondities in de in de meta-analyse geïncludeerde studies, was het niet mogelijk om het effect dat bepaalde omgevingscondities op de gedrag uitkomstmaten hebben te onderzoeken in de meta-analyse. Ik was specifiek geïnteresseerd in het moment van testen in relatie tot de licht-donker cyclus en de lichtcondities waarin de testen waren uitgevoerd. Ik zocht uit of deze twee factoren inderdaad de expressie van gedrag van het Fmr-1 KO muismodel konden beïnvloeden in sommigen van de veelgebruikte testen voor de karakterisatie van het fenotype van deze genetische muislijn. **Hoofdstuk 5** presenteert de resultaten van deze studie, die aantonen dat de tijd van testen en de lichtcondities geen effect hebben op de expressie van het gedrag in testen voor locomotie, angst en "sensory gating". Alleen in de "Acoustic Startle Response (ASR)" werd een effect gevonden, waar zowel knock-out en wildtype dieren een sterkere schrikreactie vertoonden wanneer de test werd uitgevoerd in de vroege lichtfase in vergelijking met de vroege donkerfase. Deze resultaten suggereren dat de verschillende gedragstesten verschillende sensitiviteit hebben voor de omgevingscondities van de test; transparante rapportage van deze omgevingsfactoren is dus essentieel voor het accuraat evalueren van de resultaten van een studie. Tot slot presenteert **Hoofdstuk 6** mijn ervaringen met de implementatie van het EQIPD-Quality System

(QS). Deze tool, die in detail is uitgelegd in **Appendix 1**, is ontworpen om biomedische onderzoeksafdelingen die preklinisch onderzoek doen in en buiten de academische wereld, te ondersteunen in het plannen en documenteren van onderzoeksprojecten door middel van het volgen van onderzoekspraktijken die verzekeren dat de resultaten geschikt zijn voor het voorgenomen onderzoeksdoel. **Hoofdstuk 6** had dus als doel om mijn ervaringen en geleerde lessen van het implementeren van het QS in het Kas-lab te delen om zo het gebruik van deze tool te promoten onder collega-wetenschappers en daarmee de kwaliteit van preklinisch onderzoek te verhogen.

Samengenomen, verkent het werk in dit proefschrift hoe onderzoekspraktijken zoals onder andere standaardisatie van protocollen, transparante rapportage en variatie in omgevingsfactoren, resultaten beïnvloeden en een effect hebben op hun reproduceerbaarheid en generaliseerbaarheid. Maar belangrijker, het oppert dat de replicatiecrisis een indicator is van een manier van denken in onderzoek die enigermate blind is voor datakwaliteit. Het belangrijkste resultaat van dit proefschrift is dus een appel aan de wetenschappelijke gemeenschap om de huidige onderzoekscultuur die, zoals besproken in dit proefschrift, aanzet tot suboptimale onderzoekspraktijken te veranderen. Deze verandering kan worden verwezenlijkt door het aanmoedigen van open wetenschap, preregistratie van preklinische studies, het (verplicht) gebruik van richtlijnen voor rapportage, het verbreden van criteria waarop wetenschappers worden beoordeeld en het trainen van wetenschappers in alle stadia van hun carrière in verantwoorde onderzoekspraktijken. Desalniettemin is deze lijst niet compleet en zal deze waarschijnlijk verder ontwikkelen wanneer het veld dat onderzoek doet naar de verbetering van onderzoek groeit en de aandacht en middelen krijgt die het verdient.

English summary

The ability to repeat an experiment or produce comparable results with similar methodologies has been used as a way to confirm the truthfulness of scientific findings. However, during the last decade numerous scientific reports have highlighted the low rate of the replicability of results, especially in preclinical studies. In response to this, different sources for this so-called replicability crisis have been explored. Although highly diverse, these sources are mostly linked to how the research is planned, carried out and reported; nevertheless, the drivers to incur into suboptimal research practices that lead to irreplicable results go beyond the research process and involve the incentive and reward system that supports scientific research.

Even though some sources and drivers have been identified, it is necessary to test whether different interventions within the research process can deliver more replicable results. And so the aim of this thesis was to explore different approaches to improve the way research is carried out with the final aim to increase the data quality and thus the replicability of results.

An approach to test the replicability of results it through multi-center studies as those presented in **Chapters 2 and 3**. The first, explored the level of consistency of the behavioral phenotype of the Shank2 autism-like rat model. The three sites involved had an almost fully aligned protocol, using the exact same equipment and software to score the different behaviors. The results from this study showed that protocol and equipment alignment is sufficient to generate comparable results across sites; however, the rigorous standardization of the protocol risks the generalizability of results. In contrast, **Chapter 3**, assessed the replicability of behavioral results after pharmacological interventions across seven laboratories while using different experimental designs relevant for the generalizability of results. The main difference between the different experimental designs was the degree of protocol standardization that was aligned across laboratories. In brief, this study showed that the alignment of a standardized protocol resulted in more replicable results across sites than were there was minimal standardization. Results also suggested that albeit protocol standardization, subtle variations in the experimental/environmental factors in each lab introduced variability across sites. Moreover, this variability could not be accounted by the introduction of systematic variation via diversifying the light intensity during test and the testing time-window that would increase the generalizability of results. Therefore, further studies are needed to explore more experimental designs that account for between-sites variability and/or experimental/environmental factors that ca increase variation within sites to increase the generalizability of results.

Another approach to get a grasp on replicability of results is to do a systematic review and meta-analysis of a certain intervention or phenotype of interest. In this case, literature findings can be compared against each other to assess the effect of a given variable of interest; the effect of all eligible studies are summarized in a total effect size. This approach is presented in **Chapter 4** where we evaluated the behavioral phenotype of the genetic model for Fragile-X syndrome in mice (Fmr-1 KO). The meta-analysis showed great variability of outcomes between studies for most of the behavioral categories assessed (*i.e.,* poor replicability of phenotypes across sites). Interestingly, the results from the different behavioral phenotypes assessed showed a misalignment between the symptoms shown by the mouse model and the symptoms seen in clinical practice; suggesting a re-appraisal of the limitations of this model. Moreover, the risk of bias assessment pointed out a great number of studies that do not report measures to reduce the risk of introducing bias to their study, thus the results interpretation for this model and the questionable research practices around its results has to be carefully considered. Given the underreporting of environmental conditions in he studies included in the meta-analysis, it was not possible to assess the possible influence that certain environmental conditions can exert on the behavioral outcomes evaluated in the meta-analysis. In particular, I was interested in the time of testing related to the light-dark cycle and the light conditions under which the test was carried out. Therefore, I decided to explore whether these two factors could indeed alter the behavioral expression of the Frmr-1 KO mouse model in some of the tests most often reported to assess this genetic mouse line. Thus, **Chapter 5** presents the results from this study, and these suggest that overall, the time of testing and the light conditions do not influence the behavioral expression in locomotion, anxiety and sensory gating tasks. The only effect seen was seen for the Acoustic Startle Response (ASR) task, where both knock-out and wild-type animals startled more when tested early in the light phase compared to early in the dark phase; the light condition during the test had no effect in the startle response. These results suggest that the different behavioral tasks have different sensitivity to the environmental conditions surrounding the test; thus, transparent reporting of these is critical to accurately appraise the findings of a study. Last but not least, **Chapter 6** presents my experience while implementing the EQIPD-Quality System (QS). This tool, explained in detail in **Appendix 1**, was designed to aid biomedical research units doing preclinical research in and outside academia to plan and document their research projects by following research practices that ensure the results are fit for the intended research purpose. **Chapter 6** was thus, meant to share my experience and main learnings after implementing the QS within the Kas Lab with the aim of promoting the use of this tool among colleague researchers to improve the quality of preclinical research.

Altogether, the work presented in this thesis explores how research practices such as protocol standardization, transparent reporting, environmental variation among others influence results and affect their replicability and generalizability. More importantly, it proposes the replicability crisis as an indicator of a research mindset that is somewhat blinded to data quality. Therefore, the main result of this thesis is a call to the scientific community to modify the current research culture that supports and incentivizes the suboptimal research practices discussed in this thesis. This change can be achieved with the promotion of open science practices, pre-registration of preclinical studies, the (mandatory) use of reporting guidelines, broadening the criteria to evaluate researchers and training researcher in all career stages in responsible research practices. Nonetheless, these recommendations are not exhaustive and will likely evolve as the research improvement field grows and gains the attention and resources it deserves.

Spanish summary

La posibilidad de repetir un experimento con una metodología similar y producir resultados comparables se ha usado como una manera de confirmar la veracidad de los hallazgos científicos. Sin embargo, en la ultima década, numerosos reportes científicos han indicado con frecuencia una baja replicabilidad en los resultados de estudios preclínicos. En respuesta a esto, se han explorado diferentes fuentes para la llamada "crisis de replicabilidad". Aunque altamente diversas, estas fuentes han sido mayormente ligadas a la manera en la que la investigación es planeada, ejecutada y reportada; si bien, las practicas científicas subóptimas que llevan a tener resultados no replicables van más allá del proceso de investigación, estas involucran al sistema de incentivos y refuerzos que apoyan al sistema de investigación científica.

Aunque algunas de las fuentes y causas han sido identificadas, es necesario evaluar si distintas intervenciones dentro del proceso de investigación pueden dar resultados con mayor replicabilidad. De esta manera, el objetivo de esta tesis fue explorar diferentes aproximaciones para mejorar la manera en la que se lleva a cabo investigación con el fin último de incrementar la calidad de los datos y, consecuentemente, la replicabilidad de los resultados.

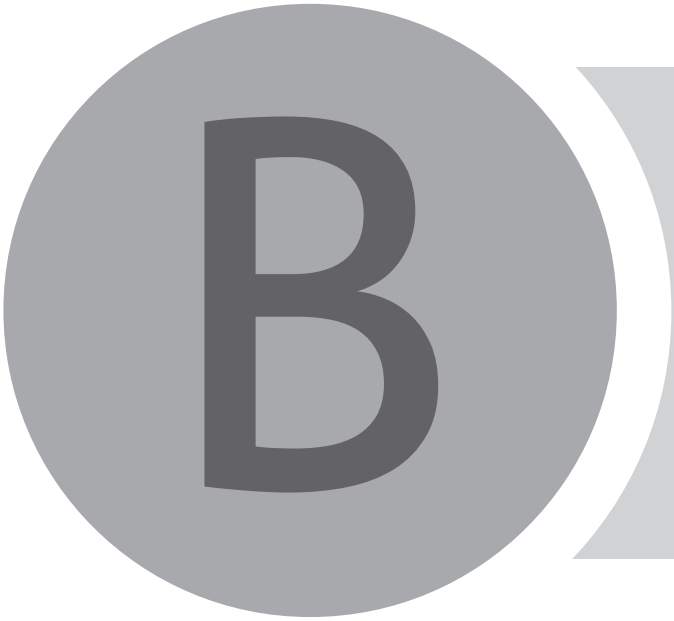
Una estrategia para examinar la replicabilidad de resultados es mediante estudios conducidos en múltiples laboratorios, como los presentados en los **Capítulos 2 y 3**. El primero explora que tan consistente es el fenotipo conductual del modelo Shank-2 de autismo en ratas entre recintos. Los tres laboratorios participantes siguieron un protocolo casi totalmente alineado, usando el mismo equipo y software para categorizar los diferentes comportamientos. Los resultados de este estudio mostraron que alinear el protocolo y equipamiento entre laboratorios es suficiente para generar resultados comparables entre sitios; sin embargo, la estandarización rigurosa de protocolos podría limitar la generalización de los resultados. En contraste, el **Capítulo 3**, hace un análisis comparativo entre 7 laboratorios que evaluaron la conducta de ratones con intervenciones farmacológicas siguiendo diferentes diseños experimentales, los cuales son relevantes para la generalización de resultados. La principal diferencia entre los diseños experimentales radicó en el grado de estandarización de protocolos intra y entre laboratorios. En breve, este estudio reveló que alinear un protocolo estandarizado entre laboratorios da resultados con mayor replicabilidad que en el escenario donde había protocolos mínimamente estandarizados. Los resultados también indican que, a pesar de la estandarización del protocolo, existen sutiles variaciones en factores ambientales/experimentales dentro de cada laboratorio que introducen variabilidad entre los recintos. Esta variabilidad entre recintos podría ser atribuida a las diferencias inherentes de los laboratorios, las cuales fueron exacerbadas por la rigurosa estandarización del protocolo. Además, esta variabilidad no pudo ser compensada por la introducción

de variaciones sistemáticas en la intensidad de la luz durante las tareas conductuales o a la ventana temporal de evaluación. Por lo tanto, se necesitan más estudios para determinar qué factores ambientales/experimentales pueden incrementar la variación intra-laboratorio para mejorar la generalización de los resultados.

Otra estrategia para comprender la replicabilidad de resultados es haciendo una revisión sistemática de la literatura y llevando a cabo un meta-análisis de cierta intervención o fenotipo de interés. En este caso, los hallazgos en la literatura pueden ser comparados entre ellos para evaluar el efecto de cierta variable de interés; el efecto de los estudios incluidos en la revisión se resume en un tamaño del efecto total. Esta estrategia es presentada en el **Capítulo 4**, donde se evaluó el fenotipo conductual del modelo genético del síndrome de la X frágil (Frm1) en ratones. El meta-análisis mostró gran variabilidad en los resultados entre estudios para la mayoría de las categorías conductuales evaluadas (*i.e.,* escasa replicabilidad del fenotipo entre estudios). Interesantemente, los resultados de los diferentes fenotipos conductuales evaluados mostraron inconsistencia con los síntomas presentados por el modelo animal y la población clínica, lo que sugiere la necesidad de reevaluar las limitaciones del modelo animal. Adicionalmente, se utilizó la evaluación del riesgo de sesgos (*Risk of bias assessment*) adoptado de estudios clínicos para evaluar el reporte de las prácticas de investigación que permiten reducir la introducción de sesgos en la ejecución del estudio (*e.g.,* estudios ciegos, aleatorización, reporte completo de todos los resultados). Este análisis indicó que un gran número de estudios reportó de manera subóptima las practicas que reducen la introducción de sesgos a su estudio, por tanto, la interpretación de los resultaos de este modelo debe de ser cautelosa. Debido a que hubo un bajo reporte de los factores ambientales en los estudios que formaron parte del metaanálisis, no fue posible evaluar la potencial influencia que ciertas condiciones ambientales pueden ejercer en los resultados conductuales evaluados en el metaanálisis. En particular, eran de interés la hora del experimento (asociada al ciclo luz-oscuridad) y las condiciones de luz en las que se llevó a cabo el experimento, debido a la importancia biológica de estos factores. Se exploró si estos dos factores podían en efecto alterar la expresión conductual del modelo Fmr1-KO en algunas de las tareas conductuales más utilizadas para evaluar el fenotipo conductual de estos ratones. El **Capítulo 5** presenta los resultados de estudios, los cuales muestran que la hora de la evaluación y las condiciones de luz no influyen la expresión conductual de tareas que miden locomoción, ansiedad y entrada sensorial (*sensory gating*). El único efecto fue para la tarea de respuesta de sobresalto acústico (*acoustic startle response*), donde los ratones knock-out y controles se sobresaltaron más cuando fueron evaluados temprano en la fase de luz que cuando fueron evaluados en las primeras horas de la fase de oscuridad. Estos resultados sugieren que las distintas tareas conductuales muestran diferente sensibilidad a las condiciones ambientales que rodean al experimento; de tal

modo que la transparencia en el reporte de estos factores escritica para ponderar de manera adecuada los hallazgos de un estudio. Finalmente, el **Capítulo 6** presenta mi experiencia al implementar el sistema de calidad EQIPD (EQIPD-QS). Esta herramienta, explicada a fondo en el **Apéndice I**, fue diseñada para asistir la investigación biomédica preclínica dentro y fuera de la academia, ayudando a planear y documentar los proyectos de investigación siguiendo practicas científicas que aseguran que los resultados sean aptos para el propósito científico destinado. De este modo, el **Capítulo 6** fue destinado a compartir mi experiencia y los principales aprendizajes después de implementar el QS en el laboratorio del Dr. Kas con el fin de promover el uso de esta herramienta entre colegas investigadores para mejorar la calidad de la investigación preclínica.

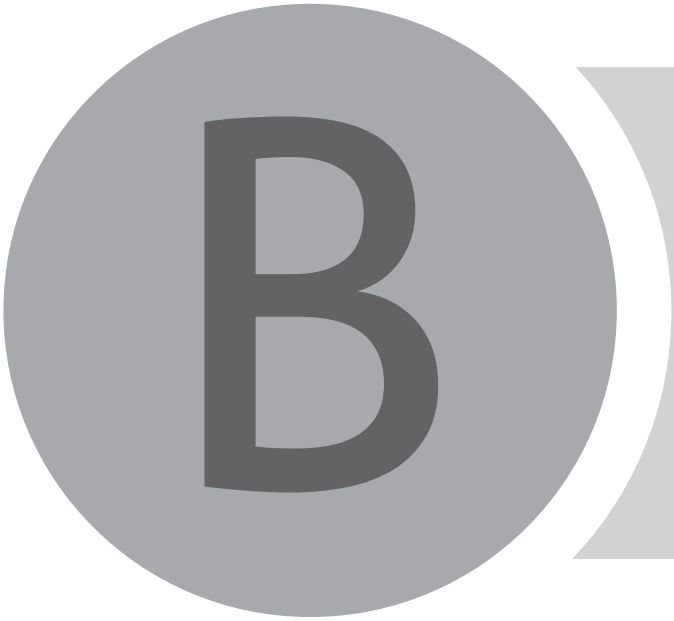
En conjunto, el trabajo presentado en esta tesis explora cómo las prácticas científicas como la estandarización de protocolos, la transparencia de los reportes, la variación ambiental, entre otros factores, influye en los resultados y afecta la replicabilidad y generalización. Importantemente, esta tesis propone la crisis de replicabilidad como un indicador de la mentalidad científica que, de una manera, está cegada a la calidad de los datos. Por consiguiente, el resultado principal de esta tesis es hacer un llamado a la comunidad científica para modificar la cultura científica actual que sostiene e incentiva practicas científicas subóptimas, como las discutidas en esta tesis. El cambio propuesto será alcanzado con la promoción de *open science*, el pre-registro de estudios preclínicos, el uso (obligatorio) de guías para reportar lineamientos de estudios, la ampliación de los criterios para evaluar investigadores y el entrenamiento en prácticas de investigación responsables en todos los niveles institucionales. Sin embargo, estas recomendaciones no son exhaustivas y probablemente evolucionarán a la par que el campo de "*research improvement"* crece y gana la atención y recursos que merece.


B

About the author

ABOUT THE AUTHOR

María was born and raised in Mexico City, Mexico where she obtained her bachelor's degree in Psychology at the National Autonomous University of Mexico (UNAM, CU). During her bachelor studies she joined the Neuropsychopharmacology and Time Estimation lab as a research assistant under the supervision of Dr. Hugo Sánchez. During 2 years, parallel to her studies, she performed pharmacological studies in rats to study the neurobiological substrates of time perception. At the end of her bachelor studies, she joined the lab of Dr. Humberto Nicolini in the National Institute of Genomic Medicine (INMEGEN, SS) where she performed her bachelor thesis, in collaboration with Dr. Vladimir Orduña from the Psychology faculty, to evaluate sustained attention after acute nicotine administration in a lesion model of schizophrenia in rats. After graduating from her Psychology bachelor with *cum laude* she moved to Groningen, The Netherlands in 2015 to join the research master program in Behavioral and Cognitive Neurosciences (BCN, c-track). During this time, María conducted her minor thesis under the supervision of Dr. Elkan Akyürek where she tested the temporal integration of colors in humans. During the second year of her master studies, she joined the lab of Dr. Martien Kas to perform her major thesis as part of the EU-AIMS consortium. This time María took part in a multi-centre study to assess a pharmacological intervention in the behavioral phenotype of the autism-like Shank2 knock-out model in rats. After completing her thesis and graduating from her master studies, María stayed in the Kas Lab to perform a PhD in the Neurobiology department of the Groningen Institute of evolutionary Life Sciences (GELIFES, RUG). The research conducted during her PhD included the exploration of early signs of tau-pathology in transgenic mice as well as the multi-centre evaluation of different experimental designs as part of the European Quality in Preclinical Neurosciences (EQIPD) consortium.


B

List of Publications

LIST OF PUBLICATIONS

Arroyo-Araujo, M., Graf, R., Maco, M., van Dam, E., Schenker, E., Drinkenburg, W., Koopmans, B., de Boer, S. F., Cullum-Doyle, M., Noldus, L. P. J. J., Loos, M., van Dommelen, W., Spooren, W., Biemans, B., Buhl, D. L., & Kas, M. J. (2019). Reproducibility via coordinated standardization: A multi-center study in a S hank 2 genetic rat model for Autism Spectrum Disorders. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-019-47981-0

Bespalov, A., Bernard, R., Gilis, A., Gerlach, B., Guillen, J., Castagne, V., Lefevre, I., Ducrey, F., Monk, L., Bongiovanni, S., Altevogt, B., **Arroyo Araujo, M**., Bikovski, L., de Bruin, N., Castaños-Vélez, E., Dityatev, A., Emmerich, C. H., Fares, R., Ferland-Beckham, C., Steckler, T. (2021). Introduction to the EQIPD quality system. *ELife*, *10*, e63294. https://doi.org/10.7554/eLife.63294

Arroyo-Araujo, M., & Kas, M. J. H. (2022). The perks of a quality system in academia. *Neuroscience Applied*, *1*, 100001. https://doi.org/10.1016/j.nsa.2022.100001

Kat, R., **Arroyo-Araujo, M**., de Vries, R. B. M., Koopmans, M. A., de Boer, S. F., & Kas, M. J. H. (2022). Translational validity and methodological underreporting in animal research: A systematic review and meta-analysis of the Fragile X syndrome (Fmr1 KO) rodent model. *Neuroscience & Biobehavioral Reviews*, *139*, 104722. https://doi.org/10.1016/j.neubiorev.2022.104722

Arroyo-Araujo M, Voelkl B, Laloux C, Novak J, Koopmans B, Waldron AM, Seiffert I, Stirling H, Aulehner K, Janhunen SK, Ramboz S, Potschka H, Holappa J, Fine T, Loos M, Boulanger B, Würbel H, Kas MJ. Systematic assessment of the replicability and generalizability of preclinical findings: Impact of protocol harmonization across laboratory sites. PLoS Biol. 2022 Nov 23;20(11):e3001886. doi: 10.1371/journal.pbio.3001886. Epub ahead of print. PMID: 36417471.


A

Acknowledgements

ACKNOWLEDGEMENTS

All the work presented in this thesis wouldn't be possible without the guidance and support of so many people through the last years; here are some of the most notable ones.

Martien, I am very grateful to you for trusting me in taking part of EQIPD. You had already supervised me during my master thesis so I guess so saw in my something that I couldn't at that time. You've always gave me the impression of seeing a step ahead of situations, I admire this of you. Thank you for giving me that little push I needed every now and then, and for being so supportive with the immigration hiccups. I am very happy to have had the opportunity of working with you and learning from your networking and marketing skills, I still need to improve mine. I must say that your guidance was nicely complemented by Peter and Robbert's. **Peter**, although in the end my project turned out to be further from your heart than expected, I appreciate your involvement. I really benefited from your deeper and critical feedback, for reminding me to not get lost in the details and be more positive about my results. **Robbert**, although you were the last to join my PhD project your contributions were equally important. I am very thankful that you were so open to discuss data with me over coffee, chocolate and other delights present in your office. Thank you as well for the support you gave me with topics beyond the lab work and for showing me the more human/realistic side of being a researcher in academia. An understanding conversation goes a long way, thank you for that!

To all the co-authors and collaborators in this book, specially to the EQIPD members, thank you for being a source of inspiration to me and for sharing your valuable knowledge. **Sylvie, Steve** and **Daniel** thank you and PGI for hosting me in New Jersey (USA) and Brno (Czech Rep.) to show me how preclinical research works outside academia.

The data collected for this thesis and the one that didn't make it to this book cannot be attributed solely to me but to the many hands that taught me and helped me through the research process. A big thanks to **Jan K., Kunja, Wanda, Jan B., Christa** and **Roy**, without you my wet-lab skill would still be zero and the so many hours processing brain samples simply would have not been possible. In addition, **Pleunie,** thank you for your dedication and friendliness towards me and my students, even when we were equally oblivious. To the animal caretakers and welfare body, **Martijn, Wendy, Roelie, Jesse and Miriam,** thank you for the many successful breeding plans and IvD discussions held through the years, I couldn't have done it without your expertise. I am also very thankful to my students: **Emelien, Rióna, Zain** and to some extent **Marthe, Jaya** and **Casper** although most of the work done within your projects is unfortunately not present in this

book, I'd like to thank you for showing interest in my project and sharing good and some more stressful times, you all did a great job!

I would also like to thank my colleagues from the **Kas Lab** for their support in and outside the lab. **Kevin** and **Betty** (now as lab grandparents) thanks for your clear guidance and support for starting the PhD together with **Renate, Filipa** and **Emma,** without you surgeries wouldn't have been so challengingly fun. **Mila, Bente, Suzanne, Iris, Remco, Tim, Celina, Magda, Freja, Michael** thanks for the shared lunch breaks, drinks, concerts, BCN events, help with experiments and passionate discussions. You definitely made these years more enjoyable and reminded me the importance of developing a network/community of like-minded and nice people to take a step out of routine and appreciate the landscape :)

Many thanks to the rest of colleagues from the Neurobiology department. **Diana** gracias por tu apoyo y consejos en el breve tiempo que coincidimos. **Adi** thank you for your endless good vibes and contagious smile. **Frank, Natalia, Francien, Iris** and **Youri** thanks for the many Km's ran through Groningen and beyond, you definitely help me to stay sane and active during the lockdowns. **Youri** thank you for the many adventures together in the lab, at home and abroad, during the lockdowns and the in-betweens. You nicely complemented the PhD time with fun time; thanks for showing me so much about the Netherlands (and Ajax) and for sharing your passionate and supportive being with me, I admire you and wish you all the best! A special thanks to **Wouter** and **Marieke** for welcoming me in your lovely home and family; spending so many gezellig dinners and nice conversations together really made me feel at home in Amsterdam. Haartelijk bedankt vor alles!

To the third-floor office colleagues, **Sjoerd, Renske, Laura, Frank, Lauren, Yifan, Anouk, Jioahue, Maarten, Kornelja, Francesca and Minqi** thanks for bringing a nice atmosphere to our quiet and hard-working bubble. Keep our traditional masterettes going, I wish y'all all the best! **Iris,** ma chérie, a special thanks to you! You've been witness of my best and worst ideas in and outside work, thanks for being an accomplice to not taking ourselves so seriously and creating a space for supporting each other. I really admire your kindness and strength and I'm very thankful for our diverse Frisian-Mexican friendship <3

A mis amigues mexas de este lado del charco, gracias por ser mi familia lejos de casa. **Andrea, Luis, César, Héctor, David, Paquito, Pam** gracias por permanecer a mi lado en las buenas y en las malas. Su apoyo, amor y compañía han hecho de estos años una experiencia más linda. Gracias por hacerme reír, bailar, platicar, beber y cantar hasta

que la vida se sintiera ligera de nuevo. Les amo! **Sarah** you are also part of my family away from home. Thank you so much for the eye-opener discussions and the woman empowering vibes. I admire your infinite strength and shared love. Theo is the luckiest to have you as his mother!

¡Una mención honorífica a mi familia que desde lejos me han dado su apoyo y me han inspirado de tantas maneras! **Clau** y **Cayo,** gracias por enseñarme tanto de la vida, y sobre cuestionar lo que es importante para mí. **Madre** mía, no sé qué haría sin tu infinito amor y entendimiento por esta vida. Gracias por enseñarme lo que significa tener la fortaleza para seguir a tu corazón y ponerle una sonrisa al mundo ante todas las adversidades. Siempre admiraré cómo tu inmensa gratitud cambia mentalidades. Eres la persona más admirable de este mundo. Te amo muchísimo, ma.

Luis, hermana mía, agradezco tu apoyo incondicional durante este viaje. Tu dedicación es admirable, gracias por compartir conmigo esa pasión que tanto te caracteriza y por combatir juntas al heteropatriarcado patriarcado cisgénero machista y opresor. Gracias por los tragos de frustración compartidos y también por los tragos de felicidad. Espero que sigamos compartiendo muchas aventuras con menos tusa y más logros por venir. Te amodoro!

Renate, I am super grateful we embarked into this adventure at the same time, you were an immense source of guidance, support and most importantly, inspiration for me. Fortunately, this turned into a super fun and lovely friendship which I cherish fondly. I am very thankful to you for the insights shared, the many questioned standards and the meaningful conversations. Thanks for making me a more environmentally responsible person, I am super proud of you (and your Spanish skills). The Dutch piece of my heart is happy to keep on celebrating our friendship and achievements in the many years to come :)

Last but not least, a big thank you to myself for going after what you want and standing up for what you believe! It's been a bumpy road but I am very proud of you, you truly made the best out of it. Keep it going!