

**Animal welfare from the animal's perspective:
Tapping into their psychological experiences**

Dissertation

for the award of the degree

“Doctor of Philosophy (Ph.D.)”

Division of Mathematics and Natural Sciences

at the Georg-August-Universität Göttingen

within the doctoral program Behavior and Cognition (BeCog)

at the Georg-August-Universität School of Science (GAUSS)

Submitted by

Lauren C. Cassidy

from California, United States of America

Göttingen, 2022

Thesis Committee

Dr. Dana Pfefferle

Welfare and Cognition Group, Cognitive Neuroscience Laboratory, German Primate Center, Leibniz Institute for Primate Research, Göttingen

Prof. Dr. Stefan Treue

Cognitive Neuroscience Laboratory, German Primate Center, Leibniz Institute for Primate Research, Göttingen

Prof. Dr. Annekathrin Schacht

Affective Neuroscience and Psychophysiology, Georg-August-Universität Göttingen

Members of the Examination Board

First referee:

Dr. Dana Pfefferle

Second referee:

Prof. Dr. Annekathrin Schacht

Further members of the Examination Board

Dr. Stefanie Keupp

Cognitive Ethology, German Primate Center, Leibniz Institute for Primate Research, Göttingen

Prof. Dr. Rabea Hinkel

Laboratory Animal Science, German Primate Center, Leibniz Institute for Primate Research, Göttingen

Prof. Dr. Alexander Gail

Sensorimotor Group, Cognitive Neuroscience Laboratory, German Primate Center, Leibniz Institute for Primate Research, Göttingen

Date of oral examination: 14 September 2022

*For Edgar, Derek, Hanjo, and the long-tailed ladies
who made this thesis possible.*

Contents

Acknowledgements	iii
Summary	1
Zusammenfassung	5
1 General Introduction	9
1.1 The concepts of animal welfare and well-being	10
1.2 Historical approaches to measuring animal welfare	11
1.3 Types of welfare parameters	12
1.4 Methods of welfare and severity assessments	12
1.5 Moving forward towards objective welfare and severity assessment	16
1.6 Refining animal welfare through literature reviews and search strategy tools	24
1.7 Animals in basic and biomedical research	25
1.8 Overview of thesis chapters	28
2 Choice-based Severity Scale (CSS): A novel concept for severity assessment in laboratory animals	29
Abstract	31
2.1 Introduction	32
2.2 Methods	36
2.3 Results	43
2.4 Discussion	46
3 Adult male rhesus macaques (<i>Macaca mulatta</i>) can associate abstract stimuli with long-delayed reinforcement	51
Abstract	53
3.1 Introduction	54
3.2 Materials and methods	56
3.3 Results	65
3.4 Discussion	74
4 The dot-probe attention bias task as a method to assess psychological wellbeing after anesthesia: A study with adult female long-tailed macaques (<i>Macaca fascicularis</i>)	79
Abstract	80
4.1 Introduction	82

4.2	Materials and methods	83
4.3	Results	84
4.4	Discussion	85
5	Comprehensive search filters for retrieving publications on nonhuman primates for literature reviews (filterNHP)	99
	Abstract	101
5.1	Introduction	102
5.2	Description	103
5.3	Example	108
5.4	Comparison and critique	109
5.5	Conclusion	110
6	General Discussion	113
6.1	Summary of chapters	113
6.2	Discussion of choice-based preference and attention bias testing	114
6.3	Thoughtfully developed search filters can enhance literature search strategies . .	123
6.4	Future outlook	124
6.5	Overall conclusions	128
7	References	131
8	Appendices	161
8.1	Appendix A: Supplementary material for Chapter 2 (Choice-based Severity Scale)	161
8.2	Appendix B: Supplementary material for Chapter 3 (Long-delay learning in non-human primates)	172
8.3	Appendix C: Supplementary material for Chapter 4 (Dot-probe attention bias task)	176
	Declaration	193
	Curriculum Vitae	195

Acknowledgements

Many thanks go to a number of people who helped me along this long journey. The completion of this thesis would not have been possible without their support. First and foremost, I would like to thank Dana Pfefferle for expertly guiding me over the course of my Ph.D. Your mentorship has greatly improved my scientific, project management, and writing capabilities to name a few. I have thoroughly enjoyed discussing our projects and strategizing solutions to the issues that inevitably came up. Your support goes beyond that of a mentor but of a friend.

I also extend my warm thanks to Stefan Treue, who welcomed me into to the lab, helped to guide and train me to think critically, provided spot-on manuscript advice, and for the many scientific opportunities I was able to learn from. Along the same lines, I am very thankful for the opportunity to collaborate with Alexander Gail on the projects within this thesis, and that he agreed to join my thesis defense evaluation committee. I am also grateful to Annekathrin Schacht for agreeing to join my thesis committee, helping along my exploits into human psychophysics, and for kindly accepting to review my thesis. Additionally, I would like to thank Stefanie Keupp and Rabea Hinkel for agreeing to be members of my thesis defense evaluation committee.

I would also like to thank and recognize the technical assistants and animal care staff for their support. Particular thanks go to Janine Kuntze for training and taking care of Bärchen in the early days of my Ph.D projects, and to Leo Burchhardt for her expert training advice and helping out when needed. Many thanks go to Klaus Heisig for building the equipment I needed for my experiments. I also thank Sina Plümer for helping to wrangle the Betriebstechnik and providing helpful support throughout my projects. Thanks also go to Luisa Guo, Ronja Mielsch, and Dr. Daniela Trinca Bertazzi Lazzarini for their additional support. Thank you to Beatrix Glaser for all the administrative help and to Matthis Drolet for technical and administrative help during his time in the lab. I am extremely grateful to Ralf Brockhausen for touching the eXperimental Behavioral Instruments (XBIs) when it “wasn’t working”, teaching me how to program cognitive tasks and extracting the data, and his collaboration on many of the projects I carried out in the last years. Thank you to the animal care and veterinary staff for taking very good care of the monkeys and for working around my testing schedule.

I have had the great opportunity to collaborate with many wonderful people over the last years. Special thanks go to the core current members of the Welfare & Cognition group for our

collaborations continuing to develop and refine the XBI as a training and enrichment device: Antonino Calapai, Pinar Yurt, Anahita Nazari, Dana, and Ralf. Cathalijn Leenaars introduced me to the world of systematic reviews and provided excellent and prompt feedback on all things related to our shared projects. Thanks go to Irene Lacal, Anahita, and Dana for the long hours spent on the systematic review investigating methodologies of preference and choice tasks in non-human primates. I also have Emily Bethell to thank for the multiple collaborations developing cognitive tasks to test aspects of non-human primate cognition. Thank you to Susann Boretius for letting us work with the lovely long-tailed ladies. Many thanks also go to Roger Mundry for teaching an excellent course in statistic and being very generous with his statistical advice. Anna Strüber and Tina Zahrie gave me the opportunity to teach and mentor two excellent students, and who helped immensely with the human psychophysics experiments projects we tackled.

Vigorous thanks go to Julia Novak, Federica Dal Pesco, Benedict Wild, and Delphine de Moor for reading early versions of this thesis. Many thanks also go to the lab members I have worked with and befriended since I started at the German Primate Center. I would like to specially thank Zurna Ahmed for her enthusiastic support and friendship, and Ole Fortmann for being my lab retreat pal and for his friendship. Outside of the lab, I have many friends to thank for their great advice, friendship, top bant, and more: Federica, Matthis, Sarah Plací, Simon Stephan, Lauriane Faraut, Delphine, Oliver Leihsa, James Stranks, Baptiste Sadoughi, and Nadine Müller. Special thanks to Johanna Prüfer for our great friendship that blossomed through running club.

Heartfelt thanks go to my family for figuring out how to cope with the long distance between us. Their support throughout this journey means the world to me and I would not have been able to take on this challenge without them. Maggie and Phili deserve special thanks for their interspecies social buffering effects and for providing me with too many cat photos to share and stories to tell.

Most of all, I thank Alan Rincon, my life partner, for his love and support in many aspects of my Ph.D and personal life. Not only did he help me bring filterNHP to life but he provided great feedback and discussion when I needed it throughout my Ph.D. Most importantly, Alan continued to remind me to care for myself and has helped me become the person I am today. I look forward to the next steps of our adventure together!

Summary

The primary goal of animal welfare science is to make an accurate assessment of an animal's well-being at a given time. To facilitate this goal, animal welfare and severity assessments were developed to integrate information from multiple domains of well-being, such as behavior, physiology, cognition, into a common framework. In these assessments, welfare parameters are typically scored each time that animal welfare is assessed. This system of scoring, however, has several issues. For instance, the different criteria within a welfare parameter are often ranked by how much they impact animal well-being based on anthropocentric judgements. The scaled difference between criteria is also often arbitrary, determined anthropocentrically, or/and not even considered. These issues result in welfare and severity assessments that lack critical insight from the animal's perspective.

The central aim of my thesis was to develop more systematic and scientific approaches that tap into the psychological experiences of animals, that can be used to reform the structure of welfare and severity assessment to be more objective and animal-centric. In this thesis, I developed and applied choice-based preference testing and tasks detecting affect-mediated changes in cognition in two non-human primate species commonly used in basic and biomedical research (Chapter 2 and Chapter 3: rhesus macaques, *Macaca mulatta*; Chapter 4: long-tailed macaques, *M. fascicularis*). In the final study (Chapter 5), I developed a scientific literature search tool that can help future comprehensive literature reviews reduce the duplication of invasive non-human primate research studies.

In Chapter 2, I proposed the Choice-based Severity Scale, a severity assessment concept that determines animals' preferences between welfare criteria that differ in their subjective value. To objectively rank and scale the options in relation to one another, I proposed that the costs of each option can be offset by providing additional reward. I tested this concept in Chapter 2 by offering monkeys a choice of where to perform a basic experimental task between two settings commonly used in systems neuroscience research (cage- or lab-based setting), and between two tasks that varied in difficulty. I found evidence that individuals differed in their subjective evaluations of the options that were compared, where one individual was more responsive to changes in reward contingencies than the other two individuals. My findings suggest that the Choice-based Severity Scale is sensitive to individual differences in subjective well-being, and that further development and refinement of the concept is warranted.

In close association to Chapter 2, I investigated the long-delay learning capabilities of adult male rhesus macaques to understand if they could learn associations between abstract stimuli and their delayed positive reinforcement in Chapter 3. Despite a delay of up to 10 minutes between selection of an abstract stimulus and the delivery of its associated reward, the monkeys reliably discovered and preferentially selected the stimulus that provided the highest reward, even when the stimuli were *novel*. Additionally, I found that the monkeys were more likely to complete trials after selecting the high reward stimulus than the low reward stimulus. My findings suggest that the monkeys retained information about the quality of the reward associated with each stimulus, and thus sustained their commitment to highly rewarded stimuli. Not only do these findings provide support to the interpretations of Chapter 2 given the use of abstract stimuli to represent the options provided, but testing for the presence and limits of long-delay capabilities offers further insight into the learning processes of animals.

I explored whether affect-mediated changes in attention bias – the tendency to attend to one type of information over another – could be detected by the dot-probe attention bias task in long-tailed macaques in Chapter 4. This task, developed for humans in cognitive psychology, measures attention biases by comparing reaction times to dot-probes replacing pairs of simultaneously presented affective stimuli (e.g., threatening and neutral faces). I showed that the task could detect attention biases to threatening over neutral affective stimuli when stimulus pairs were briefly presented during a period of putative low arousal (i.e., baseline). I found that the monkeys' attention biases deviated from this baseline pattern, by becoming avoidant of threatening stimuli, on the day immediately after experiencing prolonged anesthesia. I observed that the monkeys' baseline pattern of attention bias returned by the third day after anesthesia. Overall, my findings indicated that the dot-probe attention bias task can offer insight into the psychological well-being of non-human primates.

In Chapter 5, I created comprehensive search filters to detect scientific publications involving non-human primates and aid the development of comprehensive search strategies for non-human primate literature reviews. I found evidence that these search filters were highly sensitive to publications involving non-human primate species. I made these comprehensive non-human primate search filters freely available to other researchers through a web-based application filterNHP. Use of these search filters will enhance the quality of non-human primate literature reviews by ensuring that more topic-relevant publications are retrieved, while simultaneously reducing the time necessary for researchers to compile effective search filters.

As animal welfare scientists, it is crucial to continue to refine, develop, and validate methods of animal welfare and severity assessment. The choice-based preference and attention bias methods that I present in this thesis contribute directly to this goal and warrant further investigation as they have the capacity to address some of the long-standing issues of animal welfare and severity assessments. Combining these methodologies together and with other measures of well-being will provide further validation of these methods and insight into how

these measures align with animal welfare. Collectively, such research will guide captive animal management and research practices that will enhance and optimize animal welfare.

Zusammenfassung

Das Hauptziel der Tierschutzforschung besteht darin, eine genaue Bewertung des Wohlbefindens eines Tieres zu einem bestimmten Zeitpunkt vorzunehmen. Um dieses Ziel zu erreichen, wurden Bewertungen des Wohlbefindens und des Schweregrads von Tieren entwickelt, um Informationen aus verschiedenen Bereichen des Wohlbefindens, wie Verhalten, Physiologie und Kognition, in einen gemeinsamen Rahmen zu integrieren. Bei diesen Bewertungen werden die Tierschutzparameter in der Regel jedes Mal bewertet, wenn das Wohlbefinden der Tiere beurteilt wird. Dieses Bewertungssystem weist jedoch mehrere Probleme auf. So werden die verschiedenen Kriterien innerhalb eines Tierschutzparameters häufig danach eingestuft, wie stark sie das Wohlbefinden der Tiere auf der Grundlage anthropozentrischer Beurteilungen beeinflussen. Auch die Skalierung der Unterschiede zwischen den Kriterien ist oft willkürlich, anthropozentrisch bestimmt oder/und wird gar nicht berücksichtigt. Diese Probleme führen zu Bewertungen des Wohlergehens und der Schwere der Beeinträchtigung, denen ein kritischer Einblick aus der Perspektive des Tieres fehlt.

Das Hauptziel meiner Dissertation bestand darin, systematischere und wissenschaftlichere Ansätze zu entwickeln, die die psychologischen Erfahrungen der Tiere nutzen, um die Struktur der Bewertung von Wohlergehen und Schweregrad zu reformieren und objektiver und tierzentrierter zu gestalten. In dieser Arbeit habe ich wahlbasierte Präferenztests und Aufgaben entwickelt und angewandt, mit denen sich affektvermittelte Veränderungen der Kognition bei zwei nichtmenschlichen Primatenarten feststellen lassen, die häufig in der Grundlagen- und biomedizinischen Forschung eingesetzt werden (Kapitel 2 und 3: Rhesusaffen, *Macaca mulatta*; Kapitel 4: Langschwanzmakaken, *M. fascicularis*). In der abschließenden Studie (Kapitel 5) habe ich ein wissenschaftliches Literaturrecherchetool entwickelt, das künftigen umfassenden Literaturübersichten dabei helfen kann, die Duplizierung von invasiven Forschungsstudien an nichtmenschlichen Primaten zu reduzieren.

In Kapitel 2 schlug ich die Choice-based Severity Scale vor, ein Konzept zur Bewertung des Schweregrads, das die Präferenzen der Tiere zwischen Tierschutzkriterien ermittelt, die sich in ihrem subjektiven Wert unterscheiden. Um die Optionen objektiv in eine Rangfolge zu bringen und zu skalieren, schlug ich vor, dass die Kosten jeder Option durch die Bereitstellung einer zusätzlichen Belohnung ausgeglichen werden können. In Kapitel 2 testete ich dieses Konzept, indem ich Affen die Wahl zwischen zwei in der systemneurowissenschaftlichen Forschung

üblichen Umgebungen (Käfig oder Labor) und zwei Aufgaben mit unterschiedlichem Schwierigkeitsgrad zur Durchführung einer grundlegenden experimentellen Aufgabe anbot. Ich fand Hinweise darauf, dass sich die Individuen in ihrer subjektiven Bewertung der verglichenen Optionen unterschieden, wobei ein Individuum stärker auf Veränderungen der Belohnungskontingente reagierte als die beiden anderen Individuen. Meine Ergebnisse deuten darauf hin, dass die Choice-based Severity Scale empfindlich auf individuelle Unterschiede im subjektiven Wohlbefinden reagiert und dass eine weitere Entwicklung und Verfeinerung des Konzepts gerechtfertigt ist.

In engem Zusammenhang mit Kapitel 2 untersuchte ich in Kapitel 3 die Fähigkeit erwachsener männlicher Rhesusaffen zum Lernen mit langer Verzögerung, um herauszufinden, ob sie Assoziationen zwischen abstrakten Reizen und deren verzögerter positiver Verstärkung lernen können. Trotz einer Verzögerung von bis zu 10 Minuten zwischen der Auswahl eines abstrakten Reizes und der Aushändigung der damit verbundenen Belohnung entdeckten die Affen zuverlässig den Reiz mit der höchsten Belohnung und wählten ihn bevorzugt aus, selbst wenn die Reize *neu* waren. Außerdem stellte ich fest, dass die Affen eher Versuche abschlossen, nachdem sie den Reiz mit der höchsten Belohnung ausgewählt hatten, als den Reiz mit der niedrigsten Belohnung. Meine Ergebnisse deuten darauf hin, dass die Affen Informationen über die Qualität der Belohnung, die mit jedem Reiz verbunden ist, beibehalten haben und sich daher weiterhin für hoch belohnte Reize entschieden haben. Diese Ergebnisse untermauern nicht nur die Interpretationen aus Kapitel 2, da abstrakte Reize zur Darstellung der angebotenen Optionen verwendet werden, sondern die Prüfung auf das Vorhandensein und die Grenzen von Fähigkeiten mit langer Verzögerung bietet weitere Einblicke in die Lernprozesse von Tieren.

In Kapitel 4 habe ich untersucht, ob durch Affekt vermittelte Veränderungen in der Aufmerksamkeitsverzerrung - die Tendenz, einer Art von Information mehr Aufmerksamkeit zu schenken als einer anderen - mit Hilfe der Punkttest-Aufmerksamkeitsverzerrungsaufgabe bei Langschwanzmakaken festgestellt werden können. Diese Aufgabe, die in der kognitiven Psychologie für Menschen entwickelt wurde, misst Aufmerksamkeitsverzerrungen durch den Vergleich von Reaktionszeiten auf Punktsonden, die Paare von gleichzeitig präsentierten affektiven Reizen (z. B. bedrohliche und neutrale Gesichter) ersetzen. Ich konnte zeigen, dass die Aufgabe Aufmerksamkeitsverzerrungen für bedrohliche gegenüber neutralen affektiven Reizen aufspüren kann, wenn die Reizpaare kurz während einer Periode vermeintlich niedriger Erregung (d. h. der Grundlinie) präsentiert wurden. Ich fand heraus, dass die Aufmerksamkeit der Affen an dem Tag, der unmittelbar auf die verlängerte Narkose folgte, von diesem Grundmuster abwich, indem sie bedrohlichen Reizen aus dem Weg gingen. Ich beobachtete, dass die Affen am dritten Tag nach der Narkose wieder zum Grundmuster der Aufmerksamkeit zurückkehrten. Insgesamt deuten meine Ergebnisse darauf hin, dass die Aufmerksamkeitsaufgabe mit der Punktsonde einen Einblick in das psychologische Wohlbefinden von nichtmenschlichen Primaten geben kann.

In Kapitel 5 habe ich umfassende Suchfilter entwickelt, um wissenschaftliche Veröffentlichungen mit nichtmenschlichen Primaten zu erkennen und die Entwicklung umfassender Suchstrategien für Literaturübersichten über nichtmenschliche Primaten zu unterstützen. Ich fand Hinweise darauf, dass diese Suchfilter sehr empfindlich auf Veröffentlichungen reagieren, die nichtmenschliche Primatenarten betreffen. Ich habe diese umfassenden Suchfilter für nicht-menschliche Primaten anderen Forschern über eine webbasierte Anwendung filterNHP frei zur Verfügung gestellt. Die Verwendung dieser Suchfilter wird die Qualität von Literaturübersichten über nichtmenschliche Primaten verbessern, indem sichergestellt wird, dass mehr themenrelevante Veröffentlichungen gefunden werden, während gleichzeitig die Zeit, die Forscher für die Zusammenstellung effektiver Suchfilter benötigen, reduziert wird.

Für Wissenschaftler, die sich mit dem Tierschutz befassen, ist es von entscheidender Bedeutung, die Methoden zur Bewertung des Wohlbefindens und des Schweregrads von Tieren weiterhin zu verfeinern, zu entwickeln und zu validieren. Die in dieser Arbeit vorgestellten wahlbasierten Präferenz- und Aufmerksamkeitsmethoden tragen direkt zu diesem Ziel bei und sollten weiter untersucht werden, da sie in der Lage sind, einige der seit langem bestehenden Probleme bei der Bewertung des Wohlergehens und des Schweregrads von Tieren anzugehen. Die Kombination dieser Methoden miteinander und mit anderen Maßstäben für das Wohlbefinden wird eine weitere Validierung dieser Methoden ermöglichen und Aufschluss darüber geben, wie diese Maßstäbe mit dem Wohlergehen der Tiere in Einklang stehen. Insgesamt werden diese Forschungsarbeiten die Haltung von Tieren in Gefangenschaft und die Forschungspraktiken leiten, die das Wohlergehen der Tiere verbessern und optimieren werden.

Übersetzt von DeepL.

Chapter 1

General Introduction

Non-human animals (hereafter, ‘animals’) play a vital important role in our lives. Not only are animals our companions and sources of nutrition, but they also are a link to the natural world and its biodiversity, help us understand our own evolution, and advance applied and basic scientific knowledge (e.g., applied: gene therapies; basic: function of the brain). For captive animals, it is common public consensus in the Western world that it is our duty as the stewards and caretakers of animals to at least fulfill their basic needs (e.g., food, shelter) and that unnecessary suffering is limited (Leaman et al., 2014; Lund et al., 2012). These views are reflected in the legislature of many countries, with specific references to environmental, nutritional, and social conditions, and protection from pain, injury, and suffering (e.g., European Commission Directive 2010/63/EU; US Animal Welfare Act; UK Animal Welfare Act, German Animal Welfare Act).

Animal welfare science evolved from within veterinary medicine into a truly comprehensive discipline that has embraced many aspects of the biological sciences including ecology, neuroscience, animal behavior, genetics, cognitive science, and evolution (Dawkins, 1998; Marchant-Forde, 2015). Historically, however, the field was heavily influenced by ethology and early animal welfare scientists generally shied away from recognizing the emotion-like states (i.e., affect)¹ and subjective experiences of animals as scientifically valid (Fraser, 1999). The primary limitation to understanding these internal phenomena was (and still is) that direct observation is difficult and linguistic report is not natural in animals.

Only within the last several decades has there been a renewed interest in applying scientific methods for understanding the psychological experiences (i.e., subjective experiences, affect) of captive animals (Marchant-Forde, 2015). This resurgence has led to the development of numerous scientific approaches, ranging from preference testing (Kirkden & Pajor, 2006; Schapiro & Lambeth, 2007) and measuring affect-mediated changes in cognition (e.g., Mendl et al., 2009), to cross-species neurological studies (e.g., Panksepp, 2011). Despite these developments, captive animal welfare and severity² assessments have lagged behind in using

¹There is a lack of consensus on the definition of emotion with respects to animals (de Vere & Kuczaj, 2016), driven by the wide debate about which species have the capacity for consciousness (see Paul et al., 2020). Emotions are short-term valenced states (e.g., negative or positive) in response to external experiences that are consciously felt in humans (Mendl & Paul, 2020). I use the terms ‘emotion-like’ and ‘affect’ throughout this thesis to refer to the non-conscious components (e.g., behavior, neural activity) of emotion when referring to animals (Mendl & Paul, 2020). ‘Affect’ is the umbrella term for all valenced experiences (Howarth et al., 2021; Kremer et al., 2020), and is less frequently interpreted to imply conscious experience as it is a more technical term (Paul et al., 2020).

²‘Severity assessments’ is the common term given to assessments that evaluate the impact (e.g., pain, discomfort) of an applied procedure (e.g., for biomedical purposes).

these cognitive measures of welfare to inform their design. Ideally, the metrics by which animal welfare is evaluated should be objective and reflect some aspect of an animal's actual experience and affect rather than our own perspective.

The broad aim of this thesis is to propose and develop more objective and animal-centric methods of welfare and severity assessment that capture the subjective experiences and affective states (i.e., longer lasting valenced states: Mendl & Paul, 2020) of animals. In the next sections, I introduce the concepts behind animal welfare science and what good animal welfare is. Then I elaborate further on current welfare and severity assessments, their limitations, and suggest two concepts (choice-based preferences and affect-mediated changes in cognition) for how improvements can be made. To motivate the development of methods assessing choice-based preferences and affect-mediated changes in cognition, I briefly review each field in turn. I also outline the aims of thesis chapter(s) that are associated each method. Next, I describe the value of literature reviews and search tools to the field of animal welfare. Finally, I steer the focus to laboratory animals and my study taxon, non-human primates, before briefly overviewing the core chapters of this thesis.

1.1 The concepts of animal welfare and well-being

What is animal welfare and well-being? What does it mean to have poor and good welfare? How do we measure it? These are just a few questions that come to mind when thinking about animal welfare. Generally, animal welfare refers to how well an animal copes with its external environment (Broom, 1988), whereas well-being refers to the more specific physiological and psychological aspects of how an animal is faring. Midway through the 20th century, the Brambell Report (1965) laid a foundation for the future of animal welfare science following a rise in public concern for animal welfare³. Within this report, the authors emphasized that (farm) animals should have sufficient space to carry out five essential activities, namely to lie down, stand up, turn around, stretch, and groom their body. Webster (1994) later expanded upon those essential activities by formulating the Five Freedoms: (1) freedom from hunger and thirst; (2) freedom from discomfort; (3) freedom from pain, injury, and disease; (4) freedom to express normal behaviors; and (5) freedom from fear and distress.

Notably the Five Freedoms emphasize the absence of animal suffering (Dawkins, 1980) by the prevention of poor animal welfare, rather than defining what constitutes good animal welfare (Lewejohann et al., 2020). Good welfare is now commonly characterized by the absence of negative and the presence of positive experiences and states (Boissy et al., 2007), which captive management strategies should minimize and promote, respectively (Mellor, 2016). More recently animal welfare scientists have emphasized the importance of animals

³The publication of *Animal Machines* by Harrison (1964) shocked the British public with its descriptions of common living conditions experienced by livestock in production systems at the time. Given the commotion that the book created, the British government was compelled to act (Fraser, 2009).

experiencing a good quality of life rather than merely avoiding experiencing unfavorable conditions (Broom, 2007; Green & Mellor, 2011; Webb et al., 2019). Additionally, animal sentience – the capacity to experience subjective states (e.g., suffering, pleasure) – has been recognized as a central component of animal welfare (Broom, 2014; Browning & Veit, 2022), and is now acknowledged in the animal welfare laws of Canada, the European Union, the United Kingdom, Australia, and New Zealand (see Browning & Birch, 2022).

1.2 Historical approaches to measuring animal welfare

Historically, there were three approaches for assessing animal welfare: normal physiological function, the ability to express naturalistic behavior, or animal affect (Fraser, 2009). Monitoring an animal's physiology, such as health and bodily function, is relatively straightforward as these parameters are often discrete and measurable (Marchant-Forde, 2015). Here, the emphasis is that animals are thriving – free from disease, injury and deformation, malnutrition, abnormal behavior – and are capable of normal growth and reproduction (Bekoff & Meaney, 1998). The second approach to assessing animal welfare emphasizes the extent to which animals can express their natural behavioral repertoire, where good welfare is experienced when animals are able to fulfill their inherent nature (Marchant-Forde, 2015). The last approach emphasizes the assessment of animals' affective states, which are longer lasting valenced states that encompass emotion-like and mood-like (i.e., cumulative average of emotions over time: Nettle & Bateson, 2012; Trimmer et al., 2013) changes in behavior, physiology, and cognition (Mendl & Paul, 2020). This approach highlights that good welfare is not only characterized by being reasonably free from negative and unpleasant states such as pain, hunger, and fear, but that positive states, such as pleasure or happiness, are also experienced (Fraser, 2009).

Assessing only one of these approaches risks ignoring the importance of the others. While good health forms the basis of good animal welfare (Dawkins, 2006), it is important to keep in mind that an animal could be perfectly healthy but have poor quality of life. For example, separating an individual of a social species from conspecifics may prevent social injury and pain, but the benefits that sociality provides to this species will not be reaped (e.g., social buffering: Hennessy et al., 2009; Kikusui et al., 2006). Furthermore, this isolated individual would not have the opportunity to express its full behavioral repertoire (e.g., grooming conspecifics: Cassidy et al., 2020; Hannibal et al., 2018). Sustained isolation in such a circumstance would likely result in poor psychological well-being and potentially hamper more subtle aspects of physiology long-term (e.g., immunosuppression: Lilly et al., 1999).

Similarly, interpreting a single welfare parameter – a measurable or categorical unit of well-being – for the purposes of animal welfare assessment can be misleading as measurements or/and changes do not definitively reflect good or bad welfare (Rushen, 2000). Glucocorticoids, for example, are a common physiological measure of stress, but levels circulating in the body are not related to arousal valence (Romero & Beattie, 2021). Elevated glucocorticoids can

occur in response to stressful events (e.g., receipt of aggression: reviewed in Abbott et al., 2003), suggesting a negative impact to well-being, but high levels also promote positive social interactions (e.g., cooperation: reviewed in Raulo & Dantzer, 2018), suggesting a positive impact to well-being. Isolated interpretation of a single welfare parameter is further complicated by the influence of experience, genetics, age, physiological state, and season to name a few factors (Cook et al., 2000; Gottlieb et al., 2013a, 2015; Goymann, 2012; Moberg, 2000; Sheriff et al., 2011). Accordingly, the importance of assessing multiple welfare parameters and integrating the three welfare assessment approaches (physiology, natural behavior, affect) is now commonly acknowledged in animal welfare science (Broom, 1988; Dawkins, 1980; Mason & Mendl, 1993).

1.3 Types of welfare parameters

Welfare parameters can be broadly classified as environment- (e.g., enrichment, health and management practices, housing), individual- (e.g., behavior, health, physiology), or procedure-based (e.g., treatment- and procedure-specific, particular to severity assessments). Environment-based parameters are generally thought to be more objective and relatively easy to measure (e.g., presence or absence of cognitive enrichment), but welfare issues (e.g., effects of boredom due to lack of cognitive stimulation) may take time to accumulate to a discernible threshold (Johnsen et al., 2001; Mench, 2003). In contrast, changes in individual-based parameters can be detected quickly and immediate action can be initiated to reduce and prevent further animal suffering. Individual-based parameters, however, can be more ambiguous and inconsistent than environment-based parameters as their variation may be due to slight fluctuations in the individual or in the environment (Leach et al., 2008; Temple et al., 2013). Procedure-based parameters, while individual-based in their essence, are monitored when an applied procedure (e.g., surgery, drugs) is expected to inflict a certain degree of severity (e.g., pain, suffering, distress, lasting harm) upon an animal (European Parliament, 2022).

The diversity of welfare parameters begs the question of how they can be integrated in a meaningful way to give a true representation of an animal's welfare state. To address this question, welfare and severity assessments have been developed to combine this multi-dimensional information and draw a comprehensive picture of the welfare of an animal at a given time.

1.4 Methods of welfare and severity assessments

Existing animal welfare and severity assessments can be made using either qualitative or quantitative methods. In the next sections, I will discuss the advantages and disadvantages of

each method and elaborate more on severity assessments, before I propose potential methods that address those weaknesses.

1.4.1 Qualitative methods of animal welfare and severity assessment

Since the seminal work of King & Landau (2003) and Wemelsfelder et al. (2000), qualitative welfare assessments have risen in popularity as they are non-invasive and relatively easy to implement. Generally, these types of welfare and severity assessments rely on the expertise of people (e.g., husbandry staff, farmers) familiar with the animals or species to rate different questions (e.g., Subjective Welfare Questionnaire: Gartner & Weiss, 2013; Robinson et al., 2016; Robinson et al., 2017, 2018, 2021; Simpson et al., 2019; Weiss et al., 2011a; Weiss et al., 2011b; Detroit Zoological Society Individual Animal/Environment Welfare Assessment: Kagan et al., 2015; welfareTrack: Whitham & Wielebnowski, 2009, 2013) or conduct free-choice profiling, where terms are generated based on the animals' body language and behavior (e.g., Qualitative Behavioural Assessment: Clarke et al., 2016; Wemelsfelder et al., 2000; Wemelsfelder, 2007). In these assessments, observers interpret an animal's behavior within its environment; thus, the complete valence of an animal's affective state can be captured (Boissy et al., 2007; Turner, 2020). Although practical, the reliance on human observers is what draws the most criticism to these methodologies although observers are capable of interpreting animal behavior reliably (Meagher, 2009). Alongside the risk of anthropomorphism, potential confounds could arise due to experience (Bayne, 2012), confirmatory and/or expectation biases (Bello et al., 2014; Tuytens et al., 2014), and differences in the subjective interpretations of an animal's behavior (Meagher, 2009). These reasons often underpin the hesitancy to solely rely on qualitative welfare assessments and are generally why these methods are supplemented with quantitative methods.

1.4.2 Quantitative methods of animal welfare and severity assessment

Generally, the most comprehensive quantitative welfare and severity assessments are structured of multiple domains (e.g., behavioral, physiological, psychological, environmental) that correspond to aspects of well-being (e.g., Animal Welfare Assessment Grid in Figure 1.1: Honess & Wolfensohn, 2010; Justice et al., 2017; Wolfensohn et al., 2015; Wolfensohn et al., 2018). Such assessment approaches are commonplace within the farming industry (e.g., Welfare Quality®: Botreau et al., 2009; Welfare Quality Network, 2022) and have made their way into research-based settings (e.g., Hawkins et al., 2011; Honess & Wolfensohn, 2010; Rix et al., 2020; Wolfensohn et al., 2015) and zoos (e.g., Justice et al., 2017; Sherwen et al., 2018; Wolfensohn et al., 2018). Nested within these domains are several welfare parameters that can be environment-, individual-, or procedure-based in nature (Figure 1.1). Welfare parameters consist of several criteria that can be measured or/and categorized (i.e., scored; (Figure 1.1). In the Animal Welfare Assessment Grid, for example, welfare parameters are scored each

time a new assessment is made, and can occur when an animal experiences a distinct change in a parameter (e.g., a change in housing) or at pre-determined time points (Figure 1.1). Collectively, these assessments create a welfare continuum across the animal's life, where the visualization of the assessments facilitates monitoring and retrospective reflection on the impact that different conditions or/and procedures have on animal welfare (Figure 1.1). Score sheets are comprised of a similar structure, where welfare parameters are also generally nested within domains and consist of several criteria (e.g., Botreau et al., 2009; Hawkins et al., 2011; Morton & Griffiths, 1985; Welfare Quality Network, 2022). In both cases, welfare parameter scores are summed or averaged to create a composite score for each domain or/and the whole assessment.

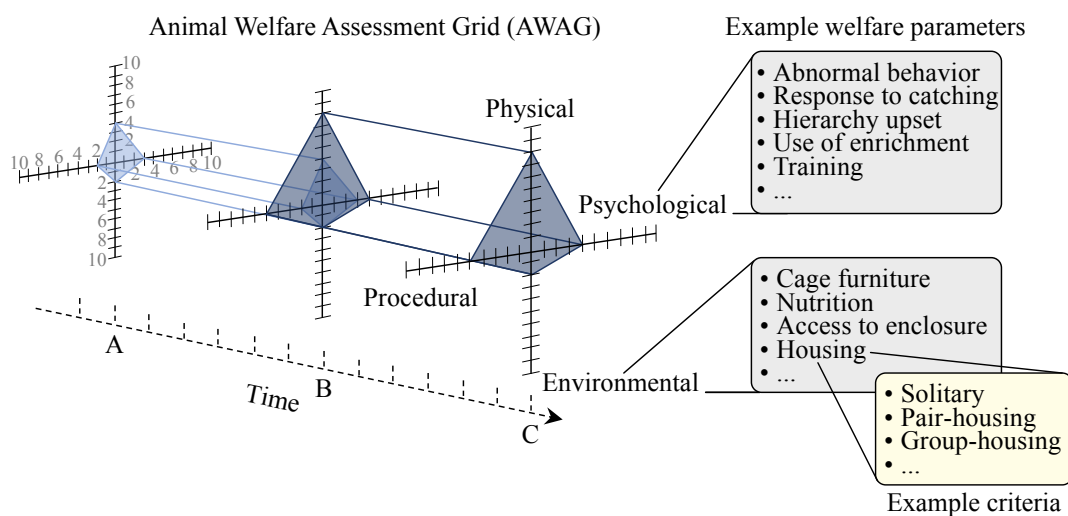


Figure 1.1: Example of the Animal Welfare Assessment Grid (AWAG), a quantitative welfare assessment. Each polyhedron represents a welfare assessment made at different points in time (A, B, C). Welfare domains are represented by the axes of each graph. Examples of welfare parameters for the psychological and environment domains are listed in the grey boxes. Example criteria of the housing welfare parameter are listed in the light yellow box. Ellipse indicate that more welfare parameters and criteria can be added than those that appear. Welfare parameter scores are compiled to create a composite score for each domain (physical, psychological, procedural, environment), which are combined to create a polyhedron (i.e., welfare assessment). The volume of each polyhedron changes depending on changes in the domain scores. Adapted from Justice et al. (2017).

It is in these welfare parameter scores that the inherent problems and limitations of these quantitative assessments become apparent. First, the rank (i.e., order) and scores (i.e., weighted difference between criteria) assigned to criteria within welfare parameters are still based on anthropocentric judgements that may not necessarily reflect an animal's subjective experience. For example, water sources in the farming industry are often scored for their cleanliness (i.e., an environment-based parameter in the environmental domain), ranging from a clean water source with fresh water (e.g., score of 0) to a dirty water source with dirty water (e.g., score of 2, Welfare Quality Network, 2022). The difference between a clean water source and a partly dirty water source with fresh water (e.g., score of 1) may actually be negligible from the animals' perspective. Consequently, the animal's perspective may not align to the ranking

and scaling of the welfare parameter dictated in the welfare assessment. Second, the level of cleanliness is still scored by human observers who may differ in their subjective evaluations of cleanliness, similar to qualitative welfare assessments. This example points to a third problem: how to compare scales across domains. Is a high score for water cleanliness perceived as severe as a high score for body condition (i.e., very thin) in the physiological domain? Arguably being thin has a more immediate and long-term impact on animal welfare, yet it is unclear whether the animals equate these conditions.

Fourth, welfare parameters are often assumed to be orthogonal. However, behavioral activity (i.e., behavioral domain) is likely highly correlated with lack of space (i.e., environmental domain), for example, which would result in the volume of the Animal Welfare Assessment Grid polyhedron to artificially balloon (Figure 1.1). Such correlated changes may lead to over- or underestimated effects on welfare, which can have strong consequential outcomes (e.g., euthanasia). Additionally, different factors can simultaneously influence animal welfare resulting in effects that are potentially compensatory or/and additive (Lewejohann et al., 2020). An animal may suffer pain from a recent injury (i.e., index of poor welfare), but compensate by positively interacting with a conspecific (i.e., index of positive welfare), resulting in little to no change in the overall welfare score, for instance. Fifth, welfare parameters are non-linear as the difference between poor welfare and moderate welfare is not necessarily the same as the difference between moderate and good welfare (Lewejohann et al., 2020). For example, the expression of natural behavior for social animals may not differ as dramatically between groups of four and five animals than the difference between one and two animals. Moreover, non-linearity stems from the fact that some parameters (e.g., physiology) do not relate to animal welfare on a one-to-one basis as one unit of change is not necessarily meaningful (Hau & Goymann, 2015; Korte et al., 2007).

1.4.3 Severity assessments

Severity assessments are conducted to evaluate the severity of an applied procedure on an animal (European Parliament, 2022). Such procedures are more typical in basic and biomedical environments, but occur in other contexts as well (e.g., tail-docking in commercial pigs, *Sus domesticus*: Nannoni et al., 2014). Severity assessments can be qualitative (e.g., dairy cows, *Bos taurus*, assessed for experimentally induced mastitis using the Qualitative Behaviour Assessment: de Boyer des Roches et al., 2018), but quantitative score sheets are more commonly used to assess additional burdens (Morton & Griffiths, 1985). These score sheets are designed to measure of pain, distress, and discomfort caused by the applied procedures (Bugnon et al., 2016; Ullmann et al., 2018). Additionally, considerations specific to the procedure are taken into account to prevent insignificant parameters being scored or missing those that are essential to assess health accurately (e.g., function of a particular body part, Bugnon et al., 2016; Hawkins et al., 2011; Rix et al., 2020).

The concerns that I brought up for quantitative welfare assessments, however, still hold for severity assessments. Procedure-based parameters are still scored by human observers, who are supposed to be trained to unequivocally recognize and score changes in animal welfare (Bugnon et al., 2016). These parameters are also not ranked nor scored in a way that is comparable across domains. Furthermore, existing score sheets may not be as well-informed as originally intended as the welfare parameters included must be appropriate for the consequences of the applied procedure. For example, body weight was the only welfare parameter of several (including behavior, general appearance, and treatment-specific parameters) included on a typical score sheet for mice (*Mus musculus*) given chemotherapy that could detect significant health deterioration (Rix et al., 2020). In contrast, body weight would be an insensitive welfare parameter for studies where tumor growth or fluid accumulation might occur as weight loss might be masked (Ullman-Culler, 1999). The animal species' natural history should also be taken into consideration. Male rhesus macaques (*Macaca mulatta*), for instance, experience natural fluctuations in their weight that are closely linked to breeding season (Bernstein et al., 1989); hence, such fluctuations should be accounted when assessing the effects of long-term studies (e.g., neuroscience experiments).

Individualized assessment is of particular importance to severity assessments. Individuals respond differently to their internal and external environments, which may be predicted by factors such as species, age, sex, and personality (Coleman, 2012; Izzo et al., 2011; Palmer et al., 2022; Sloan Wilson et al., 1994). For example, more excitable cattle exhibit greater concentrations of glucocorticoids to a stressful chute procedure and more temperamental cattle have poorer growth and immune responses than those on the opposite side of the temperament spectrum (reviewed in Burdick et al., 2011). Furthermore, each animal's life history is unique. Animals are raised differently (e.g., mother-reared versus peer-reared), they are assigned to different projects of the course of their life, they have different social experiences (e.g., dominance rank), and they are trained to do different things by different people. These are just a few examples of how different the trajectories of their lives can affect how each interprets and reacts to the outside world. Welfare and severity assessments should therefore be flexible enough to account for individual variation.

1.5 Moving forward towards objective welfare and severity assessment

It is important to recognize that the current methods of welfare and severity assessment can call attention to potential animal welfare concerns. However, the key issues of these assessments (i.e., ranking and scoring determined by humans, difficult to compare scores across domains, assumed orthogonality of welfare parameters, non-linear nature of welfare parameters, need for individualized assessment) indicate that more systematic and scientific approaches are needed to inform their structure. There are two promising concepts that we can exploit to reform the structure of welfare and severity assessments to be more objective and animal-centric. First, we

could ask the animals and let them decide. Like humans, animals learn from their experiences and develop preferences for simple (e.g., fluids: Gray et al., 2019; foods: Hobbiesiefken et al., 2021) and complex phenomena (e.g., partners: Carp et al., 2016; cognitive tasks: Ritvo & MacDonald, 2020). They are motivated to obtain resources they desire and avoid those that are aversive (Kirkden & Pajor, 2006). By offering animals choices between different criteria of welfare parameters, their decisions can be observed. From these decisions, we can infer which elements of welfare parameters are more and less preferable and consequently have a lesser or greater impact on their welfare from their perspective, respectively.

Alternatively, affect-mediated changes in cognition can offer a window into the psychological experiences of animals where the capabilities of choice-based preference testing are limited. Growing evidence indicates that animal cognition is sensitive to context (e.g., Bethell et al., 2012b; Harding et al., 2004; Nguyen et al., 2020), mental state (e.g., Burman et al., 2009; Richter et al., 2012), social experiences (e.g., Charbonneau et al., 2021; Krakenberg et al., 2020), and different stimuli (e.g., Trevarthen et al., 2019). Thus, animals' cognitive reactions to external experiences (i.e., stimuli, events, conditions, environments) that may modulate affective state could be compared to baseline measurements taken at a putatively neutral time point (Boissy et al., 2007; Crump et al., 2018). Changes in cognitive responses could indicate which external experiences enhance, compromise, or have little discernible impact on animal welfare. As there are still barriers to collecting affect-mediated cognitive measures systematically (e.g., extensive training necessary, poor reliability), exploring and testing new methods for their robustness and validity is warranted.

Together choice-based preference testing and affect-mediated changes in cognition can offer a more complete picture of how animals are actually experiencing the environments, conditions, and opportunities that we provide them. These two concepts lay the foundation for the methods that I have developed within this thesis. Over the next sections, I dive deeper into the background of choice-based preference testing and methods detecting affect-mediated changes in cognition to further set up the reasoning for their application to animal welfare and severity assessment.

1.5.1 Evaluating animals' subjective experiences using choice based-preference testing

Within the pivotal Brambell Report (1965), W.H. Thorpe (a co-author of the report) brought forward the intriguing idea of "asking" animals about their preferences for different environments given their experience with a range of living conditions (Fraser & Matthews, 1997). This suggestion led to the early development of preference tests and scientific investigation into their use for the purposes of improving animal welfare (e.g., Dawkins, 1977; Hughes & Black, 1973). The idea behind preference tests is to allow animals to judge (i.e., choose between) the alternative options provided, with the primary assumption being that the animal will act according to its own best interests (Fraser & Matthews, 1997). Here, the

term ‘preference’ denotes that a motivation to obtain or avoid one option over another exists (Kirkden & Pajor, 2006). Choice is the operational term for instances of an animals’ behavior where options are selected (e.g., approached) or unselected (e.g., avoided), such as choosing to eat an orange over an apple. These choices are guided by the animals’ subjective experiences or/and internal motivations, which can be positive (i.e., appetitive) or negative (i.e., aversive) whereby resources are approached or avoided, respectively (Kirkden & Pajor, 2006).

According to utility theory, options are comprised of multiple decision variables (i.e., costs or/and benefits) that combine into a single value (i.e., utility: Von Neumann & Morgenstern, 1944). When given a choice between options, the decider (i.e., animal) must weigh the utility of each option against the other and pick the one with the highest expected subjective utility (Walton et al., 2006). More specifically, beneficial outcomes increase utility, whereas costs decrease utility, and are dependent on the individual’s perspective. For basic options, such as food items, the decision variables are relatively straightforward as they are intrinsic (e.g., taste, nutrient content). Compound options, like experimental and husbandry procedures, are comprised of more or/and higher in complexity decision variables (e.g., transport, movement constraint, temporary social isolation) that can take a greater mental toll (e.g., discomfort or enduring isolation; see Figure 1.2) for an example comparison). Animals can combine multiple decision variables to decide if something is worth choosing or doing (see Walton et al., 2006). For example, non-human primates exhibit preferences for tasks varying in difficulty and adapt their preferences when reward or/and effort contingencies are changed (e.g., Calapai et al., 2017; Suzuki & Matsuzawa, 1997). Mice also demonstrate the ability to integrate multiple decision variables about environmental enrichment items (i.e., structural, housing, foraging) as an item’s location and time of day influences item use (Hobbiesiefken et al., 2021).

Preference tests are most often carried out through a series of binary choices, where all combinations of options are presented to the animal (e.g., preference tests in mice: Habedank et al., 2018). The frequency that each option is selected is then used to rank options relative to one another. Simple preference tests allow insight into whether a preference or motivation exists, but do not indicate anything about its strength (Duncan, 1978; Duncan, 1992). Commonly animal welfare scientists design consumer-demand type studies, where animals are presented with options that the animals have some putative motivation to obtain or avoid (Cooper, 2004; Lea, 1978). Then the difficulty required to obtain or avoid those options is systematically manipulated to determine the degree of effort (i.e., work) the animal is willing to ‘pay’ to approach or retreat from each option, respectively (Lea, 1978; Westbrook & Braver, 2015). Thus, this paid price characterizes preference or/and motivation strength, which functions as the subjective utility of that option from the animal’s perspective (Kirkden et al., 2003). Varying the physical effort (i.e., a motor action’s subjective cost or negative utility: Morel et al., 2017) required to reach and/or access a option is often used as means to determine an option’s utility. Physical effort⁴ can be measured objectively by the nature of the obstacles

⁴Although physical effort can be measured objectively, it is still interpreted subjectively and must be interpreted relative to the individual. For example, lifting weights can be more or less difficult (i.e., effortful) depending on how strong the individual is.

(typically physical or operant) placed in the way (e.g., reviewed for mice in Habedank et al., 2018). A recent consumer-demand study of mice, for example, systematically increased the number of required nose pokes to a sensor to determine how much effort the mice were willing to put in to access a putatively preferred fluid (Kahnau et al., 2022). In contrast to physical effort, understanding the mental effort (i.e., the subjective cost or negative utility of cognitive processes) needed to cope with compound options (e.g., those compromising welfare) is more difficult to quantify (Botvinick et al., 2009; Westbrook & Braver, 2015). However, the relative value of compound options can be determined by offering choices between options that differ by their decision variables (e.g., costs; Figure 1.2).. Costs may be offset by a desired decision variable (i.e., reward), which could function as a means to quantify mental effort (Figure 1.2).

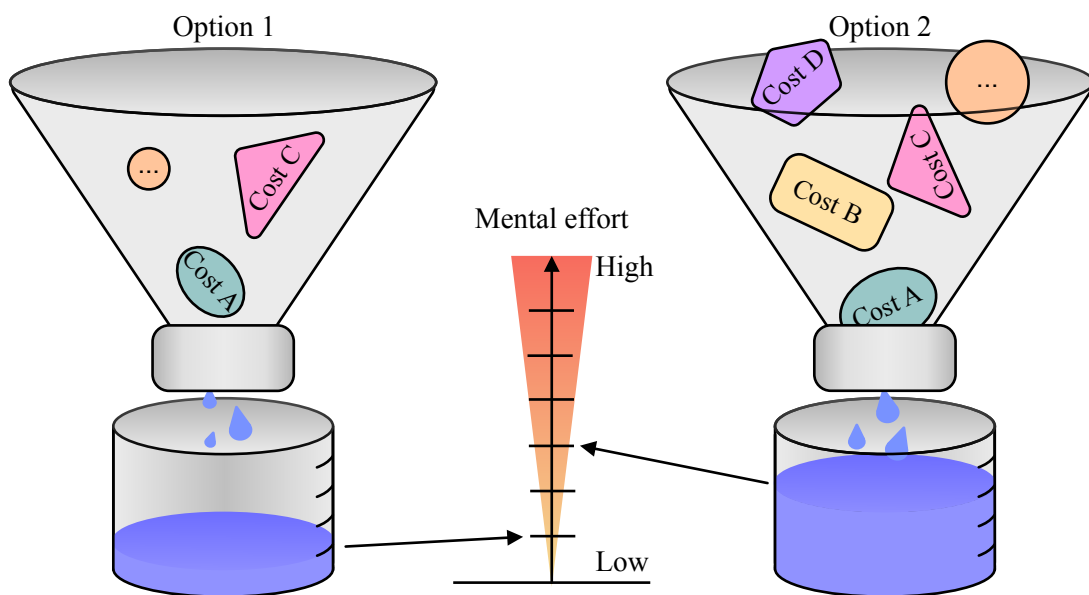


Figure 1.2: Example of two compound options and their associated decision variables (i.e., costs). Not all options have the same number or type of costs as indicated by the number of labelled shapes within each flask. Shapes with ellipse indicate that there could be more costs than those that appear in this figure. The size of the shapes are analogous to the size of the cost. The amount of reward needed to compensate for the costs of each option is indicated by the different volumes of reward in the cylinders underneath the corresponding flasks. Thus, reward amount can be used as a means to quantify the mental effort needed to cope with costs of each option.

1.5.1.1 Applying choice-based preference testing as means of welfare and severity assessment

Generally, knowledge of animal preferences can help guide and optimize management and research decisions in an objective manner (Schapiro & Lambeth, 2007). Choice-based preference testing has been used quite extensively to identify favored foods or fluids, select potential enrichment items (e.g., Hobbiesiefken et al., 2021), or determine preferred habitats to name a few (e.g., Freymann et al., 2015; review of preference tests in mice: Habedank et al., 2018). Consequently, the provision of favored items can improve individual-based welfare parameters (e.g., changes in cortisol in rhesus macaques: Arce et al., 2010; activity in lemurs,

Lemur catta, *Varecia rubra*, *Eulemur collaris*, *E. flavifrons*: Fernandez & Timberlake, 2019; anatomical features in mice: Freymann et al., 2017). But preference tests have the potential to be applied in a greater capacity with respects to animal welfare and severity assessments by asking animals which common husbandry and experimental procedures they prefer (Habedank et al., 2018; Kahnau et al., 2020). Knowledge of such preferences could allow us to objectively rank procedures in relation to one another based on the animals' perspective. Furthermore, objective scaling of these procedures can be achieved by determining how much we must pay the animals to choose each option (Kahnau et al., 2022).

The primary aim of Chapter 2 was to develop a method of welfare and severity assessment that is objective and reflects the animal's perspective. I propose that welfare parameters can be objectively ranked and scaled in relation to one another by offering animals choices between criteria and determining the amount of reward needed to pay animals to choose each component, respectively. In Chapter 2, I introduce the Choice-based Severity Scale and test this concept out by providing adult male rhesus macaques with choices between typical procedures experienced in a neuroscience laboratory. Use of the Choice-based Severity Scale will improve future welfare and severity assessments by harnessing the animal's perspective to objectively rank and scale the welfare parameters considered.

Husbandry and experimental procedures as a whole are generally abstract as they are comprised of multiple steps. Within the experiments of Chapter 2, each procedure was represented by a visual stimulus to facilitate choice-based preference testing unbiased by the environment. As one of the procedures in Chapter 2 required that the monkeys (rhesus macaques) were transported to a neuroscience setup, there was a delay (approximately 10 minutes) between their choice and its consequences (i.e., long-delay). Thus, the primary aim of Chapter 3 was to determine if adult male rhesus macaques were capable of learning long-delay associations using abstract visual stimuli. Specifically, we determined if the monkeys continued to complete trials and developed preferences for abstract stimuli despite the delivery of positive reinforcement occurring up to 10 minutes later and when stimuli were *novel*. Importantly, evidence of long-delay learning in Chapter 3 would support the interpretation of the experiments conducted in Chapter 2 and provide further insight into the long-delay learning capabilities of rhesus macaques.

1.5.2 Evaluating affect-mediated changes in cognition in animals

Providing animals with choices in every instance is often not feasible, nor is it compatible with many objectives such as those in biomedical research (Habedank et al., 2018). Nevertheless, the conditions and procedures that are experienced leave their mark on the internal affective states of animals. Affective states are principally characterized by two continuous dimensions in animals: valence and arousal (Mendl et al., 2010). Valence refers to how a stimulus is experienced—positive or negative, pleasant or unpleasant, rewarding or punishing, helpful

or harmful (Mendl & Paul, 2020). Concurrently, arousal indicates the intensity (urgency) by which a stimulus is perceived as important (Crump et al., 2020). Such states likely evolved from mechanisms developed to respond to rewarding and aversive stimuli in the environment (Cardinal et al., 2002; LeDoux, 1996). Traditionally, animal's affective states have been inferred by changes in physiology (e.g., blood pressure, heart rate, glucocorticoids: Koolhaas et al., 1999) or behavior (e.g., abnormal behavior: Novak, 2003; Novak et al., 2013; self-directed: Maestriperi et al., 1992; Troisi, 2002). However, relying on measures of physiology and behavior can be problematic as their interpretation is not always straightforward and are nontransitive to affective states (Paul et al., 2005). Thus, the development of measures sensitive to affect-mediated changes in cognition have the potential to provide critical insight into the affective states of animals and thus, psychological well-being (Harding et al., 2004; Paul et al., 2005).

Cognitive bias – the alteration of cognitive processing due to affective state – is one cognitive measure that has been proposed to offer insight into the internal states of animals (Harding et al., 2004). The premise behind cognitive bias is that an animal's affective state influences its interpretation of future events. Most cognitive bias research in animals has focused on measuring judgement biases, where animals are presented with ambiguous stimuli that they evaluate to be similar or dissimilar to previously learned (unambiguous) stimuli (reviewed in: Bethell, 2015; Clegg, 2018; Gygax, 2014; Lagisz et al., 2020; Nguyen et al., 2020; Roelofs et al., 2016). The suitability of the judgement bias task as a regularly applied welfare assessment tool, however, is primarily limited by the high risk of habituation to the ambiguous stimuli, which may result in decreased task engagement over time (e.g., Doyle et al., 2010). Thus, the development of other cognitive measures of affect is warranted (Bethell et al., 2016; Crump et al., 2018).

Tasks assessing attention bias, a class of cognitive bias, are promising alternatives to judgement bias tasks (Crump et al., 2018). Attention bias is an automatic and innate process where subjects selectively attend to one external stimulus over another (MacLeod et al., 1986). This cognitive process enables individuals to identify resources (e.g., food, mates) and dangers (e.g., predators) relevant to fitness and survival quickly (Öhman, 1986; Öhman et al., 2001). Tasks that measure attention bias exploit this innate reflex to gain insight into an individual's psychological state (Crump et al., 2018). For example, people with anxiety or depression preferentially attend to threatening (e.g., angry faces, threat words) over neutral stimuli (Barry et al., 2015; Cisler & Koster, 2010; McNally, 2019; Mogg & Bradley, 1998; Veerapa et al., 2020). Attention bias tasks are easily adaptable to animals as they include biologically relevant stimuli that trigger natural behavioral responses Bethell et al. (2019) and therefore do not require substantial training.

The dot-probe attention bias task shows promise as an animal welfare assessment tool, particularly for non-human primates (van Rooijen et al., 2017). Notably, the dot-probe task is easy to train (e.g. Kret et al., 2016) and habituation is limited as task stimuli are biologically salient and task-irrelevant (i.e., unassociated with reward). In this task, attention biases are

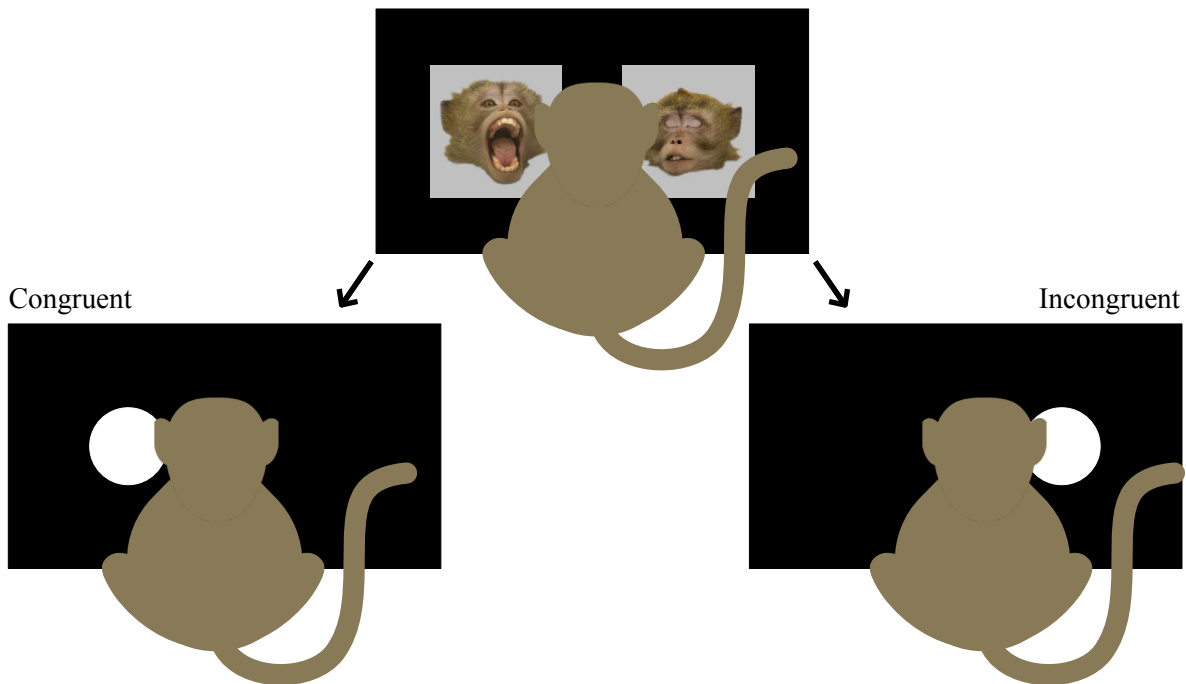


Figure 1.3: Example of the dot-probe task for non-human primates. The subjects are presented a pair of stimuli differing in valence, here aggressive and neutral faces of unknown conspecifics, for a pre-determined duration. After the stimuli disappear, a target (i.e., dot-probe) appears in the location of the aggressive face (congruent) or neutral face (incongruent). The latency to touch this dot-probe is measured and used as a proxy for attention allocation to the stimuli.

determined by differences in reaction times to a neutral target (i.e., ‘dot-probe’) appearing in the prior location of a affective stimulus (i.e., congruent) or a neutral stimulus (i.e., incongruent, Figure 1.3, Cisler et al., 2009; Cisler & Koster, 2010). Faster reactions suggest attention was already directed toward the location where the dot-probe appeared (Koster et al., 2004), whereas slower reaction times suggest that attention had to be shifted from elsewhere (e.g., from the other stimulus, Mogg & Bradley, 2005). The different processing stages of attention (initial engagement, maintenance, disengagement)⁵ can be captured by manipulating the duration stimulus pairs are presented to determine patterns of vigilance, avoidance, or both to a particular stimulus (Bradley et al., 1998; Cooper & Langton, 2006; Koster et al., 2006). Such patterns of attention processing are known to be modulated by affect. Clinically anxious people, for example, look at threatening stimuli faster and for longer durations than non-anxious people (Bar-Haim et al., 2007). Anxiety studies have also found enhanced maintenance and modulated (i.e., facilitated or impaired) disengagement to threatening stimuli (e.g., Amir et al., 2003; Fox et al., 2001; Koster et al., 2006; Rudaizky et al., 2014). Additionally, acute and chronic stressors experienced by humans have also found shifts in attention bias Shechner et al. (2012) or differences between groups experiencing different levels of stress exposure

⁵Initial engagement refers to where attention is allocated to first (Petersen & Posner, 2012; Posner, 1980), where threat-relevant stimuli are generally prioritized (Öhman et al., 2001). This initial stage is followed by maintenance or disengagement. Maintenance is when attention is sustained on a particular stimulus so incoming information is processed in more detail. Alternatively, disengagement describes the diversion (i.e., shift) of attention from the initial stimulus (Posner, 1980; Posner & Petersen, 1990).

(e.g., hunger level: Mogg et al., 1998; rocket attack: Wald et al., 2011; combat deployment: Sipos et al., 2014). Collectively these investigations in the human dot-probe literature provide the rationale for adapting and testing the dot-probe task as a means for assessing animal psychological well-being.

1.5.2.1 Applying the dot-probe task to assess the psychological well-being of animals

The dot-probe task has the potential to provide valuable insight into animal affect and well-being, but has yet to be systematically implemented in animals (reviewed in Crump et al., 2018; van Rooijen et al., 2017). Thus, the sensitivity of the dot-probe task requires validation in contexts known to elicit changes in other indices of animal welfare (Howarth et al., 2021). So far, all dot-probe studies with animals have been conducted in non-human primates (see Appendix C for a table summarizing all studies up to December 2021). Generally these studies find that non-human primates exhibit attention bias to affective stimuli over neutral stimuli, although results seem to depend on the duration stimulus pairs are presented and whether stimuli are in color or greyscale. None of these studies, however, have tested the task's sensitivity to changes in the affective state of non-human primates.

In Chapter 4 of this thesis, I tested if the dot-probe attention bias task could detect changes in the attention biases of adult female long-tailed macaques (*M. fascicularis*) following prolonged anesthesia, a context known to induce physiological stress (Lee et al., 2010; Novak et al., 2013; Whitten et al., 1998). As this study was the first of its kind, it was important to establish if the task was sensitive to affect-mediated changes in attention bias (i.e., signal of interest: Bland & Altman, 1986). This step is the first of several other criteria (e.g., repeatability, reproducibility, standardization, validation) that must be met prior to the method being generally applied to animal welfare assessment (see Howarth et al., 2021). If the dot-probe attention task proves to be robust, differences in attention bias could potentially serve as a means for ranking and scoring welfare parameters in relation to one another. These methods would be most informative for individual-based parameters, such as body condition and chronic implants (e.g., in basic and biomedical research), as comparisons between the criteria of such parameters are limited with choice-based preference testing. By developing cognitive measures of animal affect further, the accuracy of assessing animal affect will increase as well as provide much needed insight into animal psychological well-being (Harding et al., 2004; Paul et al., 2005). Only by pushing these boundaries can we determine which affect-mediated methods are sensitive, reliable, and valid.

1.6 Refining animal welfare through literature reviews and search strategy tools

Animals are studied in many different environments ranging from the wild to captivity, where different observational and experimental research approaches are applied. As a result, the scientific literature that is produced can result in substantial overlap in the topics that are investigated. Literature reviews summarize this vast and diverse literature, and highlight gaps in scientific knowledge that inform the design of future studies (Leenaars et al., 2020). In animal welfare research, literature reviews have the potential to substantially improve the scientific quality of animal experiments and generally reduce the number of animals experimented on if conducted or referenced prior to experimentation (de Vries et al., 2011).

Developing complete and comprehensive search strategies for literature reviews is time-intensive to ensure topic-relevant publications are identified so that reliable conclusions can be drawn (Leenaars et al., 2020). Such strategies are developed by thoughtfully testing of different topic-related terms to evaluate their relevance to the research topic (Hausner et al., 2012). Expertly developed topic-relevant search filters have the capacity to speed up this search strategy development process as topic-relevant terms are already collated into a formatted string. Search filters for literature involving animal research models, for example, have been developed for searches in PubMed, Embase, Web of Science, and PsycINFO, and are more sensitive to topic-relevant studies than the alternatives provided by bibliographic sources (de Vries et al., 2011, 2014; Hooijmans et al., 2010; van der Mierden et al., 2022). Generally, the use of literature search tools can help search strategies to become more standardized and reproducible (Hausner et al., 2012; Stansfield et al., 2017).

Given the importance of literature reviews to animal welfare and the greater scientific community, the primary aim of Chapter 5 was to develop comprehensive search filters for non-human primate studies. I tested the performance of these search filters against the performance of simple search strings (i.e., search strings typical of person with limited literature search experience) to validate their sensitivity. The second aim of Chapter 5 was to develop an open-source platform (called filterNHP) to share these search filters with other researchers. Future literature reviews using these filters will save time on strategy development and likely obtain more comprehensive search results. By developing comprehensive non-human primate search filters and making them accessible to others, the best practices of literature reviews will be promoted. Additionally, future invasive studies involving non-human primates may be reduced or/and refined if literature reviews using our comprehensive non-human primate filters are conducted or referenced prior to experimentation.

1.7 Animals in basic and biomedical research

Animals play an important role in the advancement of basic and biomedical research but at the expense of their suffering due to applied experimental procedures. Consequently, ethical, legal, and societal obligations expect that the severity of applied experimental procedures is minimized to the lowest extent possible in addition to standard welfare practices (e.g., close welfare monitoring, conditions that promote health and well-being, Buchanan-Smith et al., 2005; Lloyd et al., 2008). Such practices are not only important for animal welfare, but for the scientific validity of the research as poor welfare may confound results and hamper the application of any scientific findings in their translation to humans (Everds et al., 2013; Perel et al., 2007; Poole, 1997; Sneddon, 2017; Würbel, 2001). Given that animal experiments involve some degree of animal suffering, scientists generally need to justify the scientific reasoning behind the experiment by applying for specific permission from local authorities for approval prior to conducting any experiments (e.g., United States: Institutional Animal Care and Use Committee; Germany, Lower Saxony: Lower Saxony State Office for Consumer Protection and Food Safety; United Kingdom: Animals in Science Regulation Unit, Home Office). Thus, the expected consequence to animal suffering is weighed against the expected scientific benefit. The legislation and policy surrounding the use of animals for research purposes is guided by the ‘3Rs’ principles developed by Russell & Burch (1959). When feasible, animal experiments should *replace* sentient animals with non-sentient alternatives, *reduce* the number of animals involved, and *refine* experimentation by minimizing or modifying procedures and promoting welfare (Russell & Burch, 1959). But to refine animal welfare and minimize the burden that these animals bear, we need an accurate means of measuring severity. Only by understanding the extent of animal suffering during animal experimentation is it possible to make improvements that will optimize animal welfare and strengthen scientific findings in the future (Kahnau et al., 2020).

1.7.1 Non-human primates as animal research models

Non-human primates make up less than one percent of animal research models used in basic and biomedical research in the European Union (European Commission, 2020), United Kingdom (Home Office, 2019), and United States (Carbone, 2021; Service, 2019). These species are crucial for advancing knowledge across many scientific fields (e.g., translational research: Phillips et al., 2014; basic neuroscience: Capitanio & Emborg, 2008; Roelfsema & Treue, 2014; disease: Sibal & Samson, 2001; cardiovascular: Cox et al., 2017; microbiology: Kuthyar et al., 2019), however, their use for these scientific purposes is controversial and of high concern (Goodman & Check, 2002). Paradoxically, the similarities with humans that make non-human primates so useful for basic and biomedical research (e.g., lifespan, physiology, cognition, sociality) are inherently those that make them challenging to care for and ensure that their welfare needs are adequately fulfilled (Tardif et al., 2013).

Subsequently, substantial effort has gone into refining the experimental and husbandry procedures involving non-human primates (Buchanan-Smith et al., 2005; Jennings & Prescott, 2009; Prescott et al., 2021; Rennie & Buchanan-Smith, 2006a; Rennie & Buchanan-Smith, 2006b, 2006c). Such efforts have included surveying experts in the field of non-human primate welfare to determine the most useful welfare parameters for macaque species (*Macaca*, Truelove et al., 2020) and the prevalence of general management practices (e.g., environmental enrichment: Baker et al., 2007; behavioral management: Baker, 2016; social housing: Bennett, 2016). Social housing studies conducted by the National Primate Centers within the United States have also compared the effects of their housing practices and non-human primate behavior across facilities (e.g., Baker et al., 2012b, 2012a; Baker et al., 2014). These types of studies enhance the transparency of research practices involving non-human primates, can help shed light on which aspects necessitate refinement, and can make suggestions to improve the validity and reproducibility of basic and biomedical research (e.g., stable social housing: Hannibal et al., 2017; careful animal selection and conditioning: Capitanio et al., 2006). The field of neuroscience, where the practices involving non-human primates are under particular scrutiny (e.g., fluid/food restriction, movement restraint, social isolation), has made considerable strides to improve practices, for example. It is now well recognized that non-human primates with chronic implants can be socially housed without damage to implants (Roberts & Platt, 2005), refinements to fluid restriction can be made (e.g., Gray et al., 2016, 2019; Gray et al., 2017), and restraint necessary for some procedures can be trained to be voluntary (e.g., Bliss-Moreau et al., 2013; Mason et al., 2019; Ponce et al., 2016). In close association to this field, the development of automated training devices (e.g., Calapai et al., 2017; Martin et al., 2022), training protocols (e.g., Berger et al., 2018; Calapai et al., 2022), in-cage behavioral monitoring (reviewed in Knaebe et al., 2022), pose estimation (e.g., Bala et al., 2020; Mathis et al., 2018), and non-invasive eye-tracking (reviewed in Hopper et al., 2020) has exploded and allowed intensive training to be conducted without the need for restraint or social isolation.

In addition to their involvement in basic and biomedical research, non-human primates are particularly suitable for testing the new methods I have proposed to assess their subjective experiences (i.e., choice-based preference testing) and affective states (i.e., dot-probe task) in relation to common experiences in captivity. These species can be trained to conduct complex cognitive tasks due to their highly developed cognitive abilities (Roelfsema & Treue, 2014). Therefore, training necessary to learn associations between simple stimuli and more complex procedures is feasible – a necessary prerequisite for choice-based preference testing (Habadank et al., 2018). Additionally, these capabilities enable non-human primates to learn the operant responses necessary to carry out cognitive and attention bias tasks. Previous non-human primates studies have tested judgement (e.g., Bethell et al., 2012b) and attention bias tasks (e.g., Bethell et al., 2012a) in relation to typical husbandry procedures. These studies have determined that male rhesus macaques exhibit stronger changes in cognitive measures (judgement bias task: more likely to interpret ambiguous stimuli negatively; attention bias task: more avoidant of threatening stimuli) following health checks than following periods of

enrichment. These findings provide support for adapting and testing the sensitivity of other tasks to detect affect-mediated changes in cognition (Crump et al., 2018).

1.8 Overview of thesis chapters

The primary goal of animal welfare science is to continuously improve and assess animal welfare in meaningful way. Thus, accurate measures are needed to assess animal welfare, identify aspects in need improvement, and to implement and evaluate the efficacy of refinements (Browning, 2020). Recently experimental scientists have recognized the challenges of evidence-based severity assessment and have called for assessments that rely on robust and reliable welfare parameters measured across several domains (Keubler et al., 2020). I echo this call and propose that methods that incorporate the animal's perspective can reform the structure of welfare and severity assessments to be more objective and animal-centric. The studies included in this thesis support these ideas and add to the growing body of scientific literature that study animal's subjective experiences and affective states in relation to welfare.

Chapter 2 is a manuscript introducing the Choice-based Severity Scale and findings of the experiments testing this severity assessment concept in adult male rhesus macaques. This chapter is in preparation for publication.

Chapter 3 is a manuscript experimentally testing the long-delay learning capabilities of adult male rhesus macaques using abstract stimuli. The findings of this manuscript provide support for the interpretations of the previous chapter. This chapter is in preparation for publication.

Chapter 4 is a manuscript testing the sensitivity of a dot-probe attention bias task to affect-mediated changes in attention bias in adult female long-tailed macaques. This chapter is published in *European Surgical Research* as a part of a special issue on severity assessment in laboratory animals.

Chapter 5 is a manuscript describing comprehensive search filters for studies involving non-human primates available through filterNHP, an open-access web-based application. This chapter is published in the *American Journal of Primatology*.

Chapter 6 summarizes and discusses the findings of the previous chapters, and identifies several commonalities across the experimental studies. Avenues of future research with the methods developed in this thesis are also proposed.

The contributions of the individuals involved and other study-relevant information is provided at the beginning of each chapter.

Chapter 2

Choice-based Severity Scale (CSS): A novel concept for severity assessment in laboratory animals

Lauren C. Cassidy^{a,b}, Stefan Treue^{a,b,c}, Alexander Gail^{a,b,c}, and Dana Pfefferle^{a,b}

^aWelfare and Cognition Group, Cognitive Neuroscience Laboratory, German Primate Center - Leibniz Institute for Primate Research, Göttingen, Germany; ^bLeibniz-ScienceCampus Primate Cognition, Göttingen, Germany; ^cFaculty for Biology and Psychology, University of Göttingen, Göttingen, Germany

The following manuscript is in preparation for publication.

Contribution to the field

The structure of current animal welfare and severity assessments is often informed by anthropocentric judgements that may not reflect animals' perspectives. To address these concerns, the following study describes the Choice-based Severity Scale, a severity assessment concept the capabilities to rank and scale different welfare criteria using the amount of reward needed to pay animals to choose each criterion. We found evidence that individuals differed in their subjective evaluations of the options that were compared, where one individual was more responsive to changes in reward contingencies than the other two individuals. These findings suggest that the Choice-based Severity Scale is sensitive to differences in individuals and that further development and refinement of the concept is warranted. Methods like the Choice-based Severity Scale can reform the structure of animal welfare and severity assessments to be more objective and animal-centric.

Author contributions

All authors contributed to the conception and design of the study. Lauren Cassidy programmed the tasks. Alexander Gail and Stefan Treue provided the monkeys, test systems, and infrastructure support. Lauren Cassidy trained the monkeys and collected the data. Lauren Cassidy performed the statistical analyses. All authors contributed to the interpretation the data. Lauren Cassidy wrote the manuscript, with revision by Dana Pfefferle, Alexander Gail, and Stefan Treue.

Acknowledgments

We thank the German Primate Center animal care and veterinary staff for taking care of the monkeys. We also thank the technicians of the Cognitive Neuroscience Laboratory their help training the monkeys, Roger Mundry for his statistical help, and Ralf R. Brockhausen for task programming advice.

Statement of ethics

This study and the procedures involving non-human primates were conducted according to the relevant national and international laws and guidelines, including the German Animal Protection Law, the European Union Directive 2010/63/EU on the Protection of Animals used for Scientific Purposes and the Society for Neuroscience Policies on the Use of Animals and Humans in Neuroscience Research. The procedures were approved by the responsible regional government office (Niedersaechsisches Landesamt fuer Verbraucherschutz und Lebensmittelsicherheit, LAVES) under the permit number 33.19-42502-04-18/2823.

Conflict of interest statement

The authors declare that the study was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding sources

A grant from the German Research Foundation Research unit 2591 “Severity assessment in animal-based research” supported this study (<https://severity-assessment.de>), grant numbers TR 447/5-1/2, GA 1475/6-1/2, and PF 659/5-2 to Stefan Treue, Alexander Gail, and Dana Pfefferle). We also acknowledge support by the Leibniz Association through funding for the Leibniz ScienceCampus Primate Cognition (<https://www.primate-cognition.eu/en/funding-measures/audacity-funds.html>), grant number LSC-2017-01SF to Dana Pfefferle.

Supplementary material and data

See Appendix A for the associated supplementary material. The data supporting this article will become available upon publication. Code for the cognitive tasks can be provided on request.

Abstract

One primary goal of laboratory animal welfare science is to provide a comprehensive severity assessment of the experimental and husbandry conditions these animals experience. The severity of these conditions are typically scored based on anthropocentric assumptions. We propose to (a) assess an animal's subjective experience of condition severity, and (b) not only rank but scale different conditions in relation to one another using choice-based preference testing. The Choice-based Severity Scale (CSS) utilizes animals' relative preferences for different conditions, which are compared by how much reward is needed to outweigh the perceived severity of a given condition. Thus, this animal-centric approach provides a common scale for these conditions based on the animal's perspective. To assess and test the CSS concept, we offered rhesus macaques (*Macaca mulatta*) choices between two conditions: performing a cognitive task in a typical neuroscience laboratory setup (lab condition) versus the monkey's home environment (cage condition). Our data show a shift in one individual's preference for the cage condition to the lab condition when we changed the type of reward provided in the task. Two additional monkeys strongly preferred the cage condition over the lab condition, irrespective of reward amount and type. We tested the CSS concept further by showing that monkeys' choices between tasks varying in trial duration can be influenced by the amount of reward provided. Altogether, the CSS concept is built upon lab animals' subjective experiences and has the potential to de-anthropomorphize severity assessments, refine experimental protocols, and provide a common framework to assess animal welfare across domains.

Keywords: preference, choice, animal welfare, severity assessment, subjective experiences

2.1 Introduction

Animal research models (i.e., lab animals) are crucial for advancing scientific knowledge across many fields (Azkona & Sanchez-Pernaute, 2022; Bale et al., 2019; Homberg et al., 2021; Kiros et al., 2012; Meyerholz et al., 2020; Roelfsema & Treue, 2014). Good animal welfare is not only important for the health and well-being of these animals but also to the quality and validity of the research they are involved in (Jennings & Prescott, 2009; Poole, 1997). Therefore, it is our duty as researchers and caretakers of lab animals to ensure their care and welfare meets a high standard. However, animals cannot naturally linguistically report how they are experiencing different experimental and husbandry events. Caretakers and researchers must instead indirectly infer animal welfare by changes and/or differences in their physiology, natural behavior, and psychology in relation to these events. But how can animal welfare be measured objectively?

Welfare and severity assessments have been developed to quantify and understand the impact that research has on lab animals (e.g., Extended Welfare Assessment Grid: Honess & Wolfensohn, 2010; Wolfensohn et al., 2015; Qualitative Behavioural Assessment: Wemelsfelder, 2007; score sheets: Bugnon et al., 2016; Ullmann et al., 2018). In some assessments, different welfare parameters are nested within overarching domains (e.g., physical, psychological, procedural, environmental), which are broken down into the different putative conditions (e.g., procedures, events, states) that can be experienced. For example, social housing, a common welfare parameter in non-human primates, would fall under the environmental domain and could consist of four different conditions: group-housing, continuous pair-housing, intermittent pair-housing, and single housing (Hannibal et al., 2017). Each housing condition is comprised of different elements such as the number of social partners available, the duration and/or extent that physical contact with a social partner is possible, and the amount of available cage space. The state of these elements can differ between conditions; for example, the extent that physical contact with a social partner is possible is full-time in the continuous pair-housing condition and non-existent in single housing condition. Based on these differences, conditions are ranked in relation to one another and given a score based on their putative impact on welfare as assessed by humans.

During a severity assessment, welfare parameters are scored given the conditions that the animal is experiencing. Generally, these scores are combined to create a composite score for the domain. While this type of severity assessment provides a great overview of what the animal experiences over the course of its life, the hierarchy of the conditions within some welfare parameters are still determined by anthropocentric judgements. These judgements are prone to observer and confirmatory biases that may not reflect an individual's actual experience as they likely experience procedures differently (Bello et al., 2014; Tuytens et al., 2014). Presently, scores given to welfare parameters are also assumed to be comparable within and across domains. However, it is unknown whether, for example, the highest score of a welfare parameter in the experimental domain (e.g., performing a task in a laboratory setup) is equivalent to the highest score of a welfare parameter in the environmental domain (e.g., single

housing). Such comparisons are difficult as welfare parameters differ in their function and conditions are often comprised of elements that have different associated costs and benefits (i.e., ‘comparing apples with oranges’). It may also be that the different domains are not orthogonal as is often assumed as there may be dependencies between welfare parameters. For instance, the weight of an animal (e.g., clinical status) likely correlates with its daily activity (e.g., behavior).

Determining animals’ preferences can reveal how valuable certain resources are in relation to one another (Hosey et al., 1999; Kahnau et al., 2022). Often preference tests are conducted by presenting an animal with a series of binary choices among an array of options to see how frequently each option is selected in relation to the others (Habedank et al., 2018). Preference testing becomes more challenging when the options are more complex and/or abstract (compared to, e.g., choosing between favored foods or fluids: Hansell et al., 2020; Huskisson et al., 2020) as the decider, the animal, must weigh the combined costs and benefits of each. In such multi-faceted options, multiple decision variables are evaluated to optimize reward and effort and combined into a single value, the utility, which characterizes the desirability of each choice (see utility theory: Von Neumann & Morgenstern, 1944). For example, animals have preferences for tasks varying in difficulty and respond accordingly when the reward and/or effort contingencies are adjusted (Calapai et al., 2017; Suzuki & Matsuzawa, 1997). Outside of experimental tasks, animals have exhibited preferences for more complex options with respects to positive reinforcement training (e.g., Schapiro & Lambeth, 2007), environmental parameters (e.g., supplementary light: Buchanan-Smith & Badihi, 2012), environmental enrichment (e.g., Hobbiesiefken et al., 2021), and even determining their own medical treatment (Magden et al., 2013, 2016; e.g., Webb et al., 2018). Previous work has advocated for the use of preference testing to guide animal welfare assessment, particularly for determining the value of different environment-based items to animals (Habedank et al., 2018; Kahnau et al., 2020, 2022). Presently offering lab animals choices between other welfare conditions, such as experimental procedures and husbandry practices, has not been conducted to our knowledge.

Here, we propose the Choice-based Severity Scale (CSS), a novel concept for welfare and severity assessment in lab animals (see Figure 2.1). By using choice tests, we can determine which welfare conditions are preferred by animals, thus reflecting how they perceive these conditions to impact their lives. Hence, preferences can be used to rank welfare conditions as having the lowest (most preferred procedure) to highest (least preferred procedure) impact on the well-being of lab animals. Furthermore, individuals likely differ in how strongly they prefer one condition over the other and the strength of these preferences could be used to objectively scale these conditions in relation to one another. We propose that preference strength can be determined by how much is needed to “pay” the animal to choose each condition by adjusting the reward parameters (amount and/or type of fluid) experienced in each condition. Here, the difference in the reward parameters would serve as a means for objectively ranking and scaling welfare conditions in relation to one another based on the animal’s perspective.

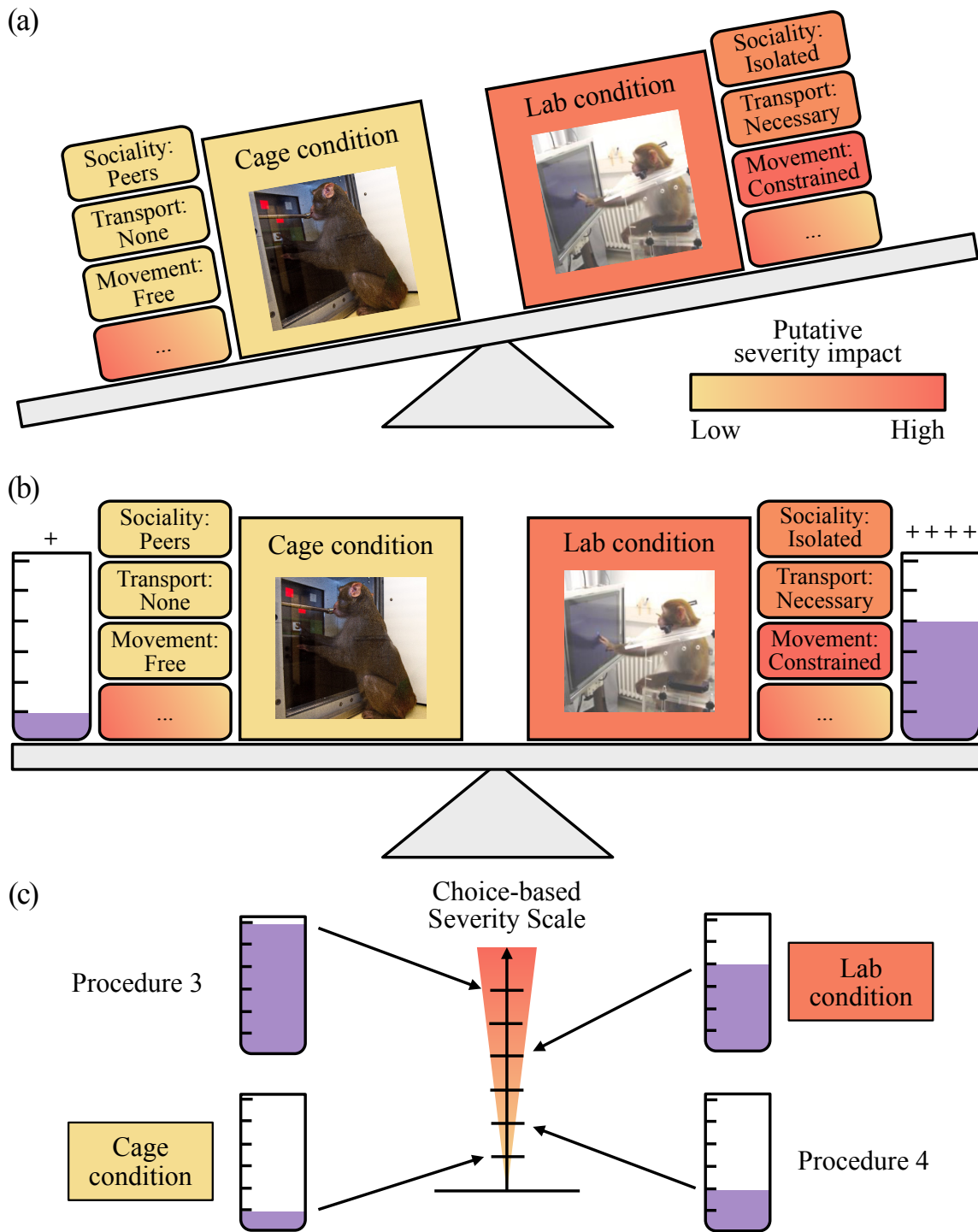



Figure 2.1: Choice-based Severity Scale: A novel concept for welfare assessment using choice-based preference testing in laboratory animals. (a) An example of two conditions, here experimental procedures, and their associated putative welfare costs and benefits (elements) positioned adjacently. Collectively the elements of the putatively less desirable condition, performing a task in a laboratory setup (i.e., lab condition), have a higher severity impact than those of the preferred condition, performing a task close to the home cage (i.e., cage condition). The ellipse indicates that there may more elements to these conditions than those we have visualized. (b) By providing the animals substantially more reward to choose the less desirable condition, we can balance the experienced severity of the two conditions in relation to one another. (c) The amount of reward needed to pay the animals to choose each condition can be used to objectively rank and scale several conditions in relation to one another on a severity scale. Ingo Bulla photographed the image of a monkey using a cognitive testing system in the cage condition.

We tested our CSS concept using non-human primates in the context of neuroscience research as our model organism and research environment. Non-human primates are particularly important animal research models in neuroscience research due to their highly developed cognitive abilities that enable them to learn new associations and perform complex sensory discrimination and motor tasks (Roelfsema & Treue, 2014). We offered three adult male rhesus macaques (*Macaca mulatta*) choices between two conditions: a typical neuroscience laboratory setup (i.e., lab condition) versus performing the same cognitive task in the monkey's home environment (i.e., cage conditions: either in the upper or lower cage of a testing compartment). The elements of these conditions differ in their states. In the lab condition, for example, the movement (i.e., movement element) of the monkey is constrained by a chair to prevent equipment from being tampered and ensure safe experimentation. In contrast to the lab condition, movement is less constrained in the cage condition as the monkey can move freely around in the cage. See Figure 2.1 and Table 2.1 for an overview of the elements of these conditions. Given the differences between these conditions, the comparison between the lab and two different cage conditions offered a robust and practical test of our CSS concept in situ (i.e., Choice-based Severity Assessment). Accordingly, we expected that the monkeys would prefer to perform a basic cognitive task in the cage conditions over the lab condition. To scale these conditions in relation to one another, we adjusted the number of fluid reward drops provided per correct trial of a basic experimental task and the type of reward in each condition. We tested our CSS concept further by offering the monkeys choices between two experimental tasks varying in trial duration, where the amount of reward provided substantially differed (i.e., CSS test). A Choice-based Severity Scale will objectively determine which aspects of research have the highest impact on laboratory animal well-being from their perspective. We provide guidelines to help implement a Choice-based Severity Scale with other species in other experimental settings.

Table 2.1: The elements (location, transport, movement, sociality) expected to differ for each condition tested and their putative severity impact.

Setting	Elements				Severity impact
	Location	Transport	Movement constraints	Sociality constraints	
upper quad.	upper quad. of testing compartment, adjacent to home cage	no	free to move within the limits of upper quad.	visual, auditory, olfactory, but no tactile contact to conspecifics	
lower quad.	lower quad. of testing compartment, adjacent to home cage	no	free to move within the limits of lower quad.	auditory, olfactory, but no visual and tactile contact to conspecifics	
lab	neuroscience setup in isolated room	yes	in primate chair	no contact to other conspecifics (isolated)	

quad.: quadrant of the testing compartment.

2.2 Methods

2.2.1 Study subjects and housing facility

We conducted the study on three adult male rhesus macaques (*Macaca mulatta*; 7, 7, and 16 years old at time of study enrollment) living at the German Primate Center, Goettingen, Germany. These monkeys were housed in isosexual pairs, with visual and auditory contact to other macaque groups. Housing was enriched and exceeded the size requirements for macaques set by EU directive 2010/63/EU (described in Cassidy et al., 2021). On days where the monkeys were not tested, they had access to monkey chow, fresh fruits and vegetables, and water ad libitum. On training and test days, fluid could be consumed by participating in the study's cognitive tasks, as is typical of neuroscience research laboratories (described in Pfefferle et al., 2018). Additionally, the monkeys were weighed each training and test day to check that their weight remained stable. Daily health monitoring of the monkeys was also carried out by veterinarians, monkey facility staff, and researchers who all have specialized training for working with non-human primates.

All monkeys had extensive training to facilitate handling for husbandry and experimental purposes, particularly for cooperatively entering and sitting in a non-human primate chair for long periods of time (Bliss-Moreau et al., 2013; Mason et al., 2019; Ponce et al., 2016). Furthermore, all monkeys had extensive experience with the basic experimental task offered in a typical neuroscience research setup (i.e., lab condition) and close to their home cage (i.e., cage condition).

2.2.2 Experimental testing apparatuses and software

We used multiple cognitive testing systems (i.e., eXperimental Behavioral Instruments: Berger et al., 2018; Calapai et al., 2017) to present condition stimuli during the CSS test and administer the cage condition tasks (i.e., basic experimental task, delivery of 2 ml bolus). These standalone systems were developed within the lab (Cognitive Neuroscience Laboratory, German Primate Center) to facilitate cage-side cognitive task training and testing. In our study, the monkeys could engage with a task by using the touchscreen and sensors equipped to the cognitive testing systems. When needed, fluid reward was dispensed via a tube positioned about 45 cm in front of the touchscreen (30.4 cm by 22.7 cm; 60-75 Hz framerate). The positioning of the reward tube on these cognitive testing systems encourages monkeys to adopt stereotypical postures when engaging with cognitive tasks (Calapai et al., 2017). Multiple cognitive testing systems were mounted to the flexible testing compartment adjacent to the monkeys' home cage. We programmed all cognitive tasks using MWorks (versions 0.8 to 0.10; <https://mworks.github.io/>). MWorks is an open-source C++-based software that allows for the design and implementation of real-time controlled behavioral tasks (Calapai et al., 2017).

2.2.3 Choice-based Severity Assessment

Our main aim was to test the CSS concept through a Choice-based Severity Assessment. We developed an experimental setup (Figure 2.2) to offer three adult male rhesus macaques a choice between performing a basic experimental task in the cage or lab condition. Choice testing was conducted in a testing compartment adjacent to the monkeys' home cage, where the choice between the conditions was presented using visual stimuli on a neutral cognitive testing system (Berger et al., 2018; Calapai et al., 2017) and the conditions were positioned on different quadrants of the testing compartment (Figure 2.2). We found that this experimental setup limited the potential influence from the environment and/or experimenter best through a series of pilot experiments summarized in the section 'Supplementary experiments' of Appendix A.

2.2.3.1 Experimental conditions of the Choice-based Severity Assessment

In our typical neuroscience research setup (i.e., lab condition), monkeys were transported in a non-human primate chair by a researcher to a small, darkened experimental room. This experimental room was equipped with devices typical of a visual neuroscience laboratory: computer monitor for the presentation of visual stimuli, various non-human primate chair attachments (e.g., sensor response box, reward delivery tube), eye-tracker, and a fluid delivery system (perisaltic pump). The researcher could administer and control cognitive tasks from a control center located just outside the door of the experimental room. In the lab condition, the monkeys were seated approximately 57 cm away from the computer monitor (59.7 cm by 33.6 cm; 120 Hz framerate). In our study, monkeys could respond to the basic experimental task presented on the computer monitor via a proximity sensor (i.e., 'sensor') and received fluid for correct trials via a reward tube attached to the non-human primate chair.

The cage conditions took place in a flexible compartment that could be divided into quadrants (approximately 80 cm by 75 cm by 90 cm) and was attached to the monkeys' home cage (Figure 2.2). All monkeys were trained to voluntarily enter this compartment for training, testing, temporary separation, experimental, and veterinary procedures as necessary. Quadrants were separated by movable sliding panels, which could be opened to shift monkeys between compartment quadrants and secured when the monkeys were present for longer durations. In the cage conditions, the monkeys could move around without restraint and had visual, acoustic, and/or olfactory contact to pair mates and adjacent social groups. Each quadrant had the capability to be equipped with cognitive testing system so that the monkeys could engage with a cognitive task without direct oversight from a researcher. Due to the cage location and ability to freely move around, we expected that the monkeys may prefer to perform their basic experimental task in this cage condition over the lab condition.

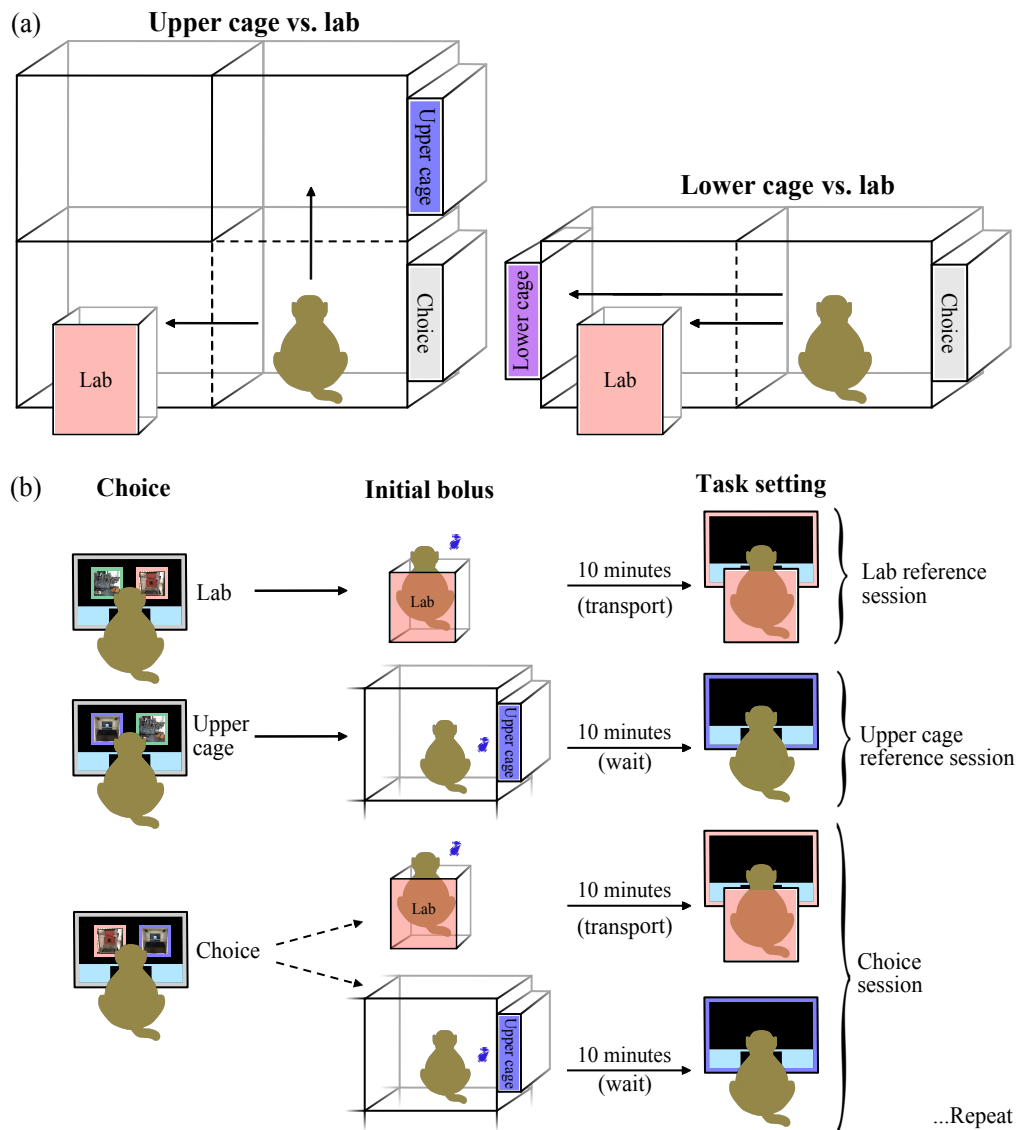


Figure 2.2: Experimental setup and study design (Choice-based Severity Assessment protocol). (a) The grey box labeled ‘Choice’ indicates the location of the neutral cognitive testing system, where the monkeys made a choice between visual stimuli representing the cage and lab conditions (i.e., condition stimuli). The cage conditions were positioned either on the upper right (i.e., upper cage: blue box, representing a cognitive testing system, labeled ‘Upper’) or lower left quadrant (i.e., lower cage: purple box, representing a cognitive testing system, labeled ‘Lower’) of a testing compartment adjacent to the monkeys’ home cage. The lab condition was positioned on the lower left quadrant (pink non-human primate chair labeled ‘Lab’). (b) Visual representation of the Choice-based Severity Assessment protocol, where the monkeys were given a reference trial for each condition prior to the choice between a cage condition and the lab condition. During reference sessions, a condition stimulus (in pink or blue) was presented simultaneously with a timeout stimulus (in green) that was unrewarded. In this example, the monkey is given two reference sessions and a choice session between performing a basic experimental task in the upper cage (blue) or lab condition (pink). Each session, the monkey selected a condition stimulus (‘Choice’), followed by a small motivational reward (‘Initial bolus’) once it was seated in the non-human primate chair (representing the lab condition) or from the cage condition cognitive testing system. The basic experimental task was started once the monkey had been transported to the neuroscience setup (approximately 10 minutes) or after 10 minutes on the cage condition cognitive testing system, matching the time course of the lab condition (‘Task setting’). Then, the monkey could conduct as many trials as desired within two hours before it was returned home. This protocol was repeated after two reference sessions and one choice session.

2.2.4 Choice-based Severity Assessment protocol

To generate a practically applicable CSS, we developed a protocol (2 reference sessions + 1 choice session; Figure 2.2) for the Choice-based Severity Assessment that allowed the monkeys to experience the full consequences of each condition (i.e., reference sessions) prior to choosing between the two conditions (i.e., choice session). Monkeys were either given one reference or choice session a day, followed by the procedure of the corresponding condition. In each condition, the monkeys could work on a basic experimental task for as many trials as they desired within two hours if they continued to engage in the task (there was a regulatory requirement to provide ample time for the monkeys to collect as much reward as they desired). The basic experimental task automatically stopped once it detected no engagement for a predefined duration (conclusion criteria was individualized), and the experimenter returned the monkeys to their home cage soon afterwards. All training and basic experimental task details are described in depth in the section ‘Basic experimental task training for the Choice-based Severity Assessment’ of Appendix A.

To scale these conditions in relation to one another, we sought to influence the monkeys’ choices by changing the reward contingencies of the basic experimental task itself. Here, we adjusted the number of fluid drops (approximately 0.3 ml each) provided per correct trial in each condition depending on the monkeys’ choices until the combination of the conditions and their corresponding amount of reward is perceived as equal (i.e., oscillating around a point of subjective equality). This adaptive approach is a popular method used in human psychophysics experiments to determine perceptual thresholds (Kingdom & Prins, 2010; Leek, 2001) and forms the basis automated training protocols to shape complex behaviors in non-human primates (Berger et al., 2018; Calapai et al., 2022). At the beginning of testing, we set the difference in reward per trial between the two conditions to be large, where the number of drops of reward per trial in the lab condition was nine times larger than in the cage condition (lab: 9 drops; cage: 1 drop). The monkeys’ preferences were assessed after every three choice sessions (i.e., bouts) and the reward per trial was adjusted so that the reward per trial of the preferred condition was reduced by 2 drops and the non-preferred condition increased by 2 drops (bounded by 1 and 9 drops). For example, if the monkey exhibited a preference for the lab condition in the first three choice sessions, then the reward per trial for that condition would be reduced from 9 to 7 drops and the reward per trial for the cage condition would be increased from 1 to 3 drops. We concluded testing if the reward per trial difference was at the extremes (1 and 9 drops) and if the monkeys made the same choice for six consecutive sessions, irrespective if a bout was finished.

2.2.4.1 Testing phases of the Choice-based Severity Assessment

Through the Choice-based Severity Assessment, we tested the monkeys over three phases where we controlled the position of the cage condition and adjusted the type of reward provided in the

basic experimental task of each condition. During the first phase, the cage condition (indicated by a cognitive testing system) was positioned on the upper right quadrant and the lab condition (indicated by a non-human primate chair) on the lower left quadrant so that the distance between a neutral cognitive testing system and each option was roughly the same (Figure 2.2). To ensure that the monkeys made choices based on a preference for the condition instead a preferred quadrant of the testing compartment, we moved the cage condition to the same quadrant that the lab condition was positioned (lower quadrant to the left of the neutral cognitive testing system). Then we tested the monkeys again in a second phase of the experiment. The type of reward per trial was the same (grape juice) for all monkeys and conditions during the first and second phases. To test if the monkeys would change their preference due to the type of reward provided in each condition, we ran a third phase. During the third phase, water was received from the basic experimental task in the cage condition (preferred option during phase 2) and the monkeys' preferred juice was received in the lab condition (see the section 'Fluid preference test' in Appendix A for more information).

2.2.4.2 Procedure for the Choice-based Severity Assessment

Each day the monkey was brought into the test compartment where the neutral cognitive testing system was mounted (Figure 2.2). Once the monkey was seated in front of the neutral cognitive testing system, the experimenter remotely triggered the start button to appear. After the monkey touched the start button, two stimuli appeared for the monkey to choose between (reference session: condition stimulus and timeout stimulus; choice session: two different condition stimuli). If a condition stimulus was touched, the experimenter opened the corresponding compartment and the monkey received a small motivational reward (i.e., 2 ml bolus of water) either by triggering the cage condition cognitive testing system (cage condition) or from the experimenter once seated in the non-human primate chair (lab condition). For lab condition choices, the experimenter then transported the monkey in the non-human primate chair to the neuroscience setup, attached the fluid reward system, and began the basic experimental task (approximately 10 minutes). For cage condition choices, the experimenter then removed the non-human primate chair and left the room for 10 min, to match the time course of the lab condition, before starting the basic experimental task. For both conditions, the monkey was returned to his home cage as soon as he stopped engaging in the task in either condition (see the section 'Basic experimental task training for the Choice-based Severity Assessment' in Appendix A for the conclusion criteria). The monkeys always made a choice and never chose the timeout stimulus during the reference sessions of the three experimental phases.

2.2.5 Choice-based Severity Scale test

To test the scaling aspect of our CSS concept further, we conducted an additional experiment in which we offered the monkeys choices between experimental tasks varying in trial duration. We

adapted the monkeys' basic experimental task to create two tasks that differed by a factor of 10 in the duration of how long the monkeys needed to hold a sensor until a stimulus change occurred (i.e., different effort needed to complete each task) and the reward the monkeys received (Figure 2.3). To determine if we could influence the monkeys' choices using reward, we provided 15 ml more reward for selecting and successfully completing the long hold task over the short hold task. To account for position biases (see the section 'Supplementary experiments' in the supplementary material for a pilot experiment), we set the position of the task stimuli to be deterministic, where the long hold task appeared in the opposite position as the last choice trial. Therefore, the task stimuli would alternate every trial if the monkey exclusively chose the long hold task. Conversely, if the monkey chose the short hold task, the position of the stimuli would remain the same.

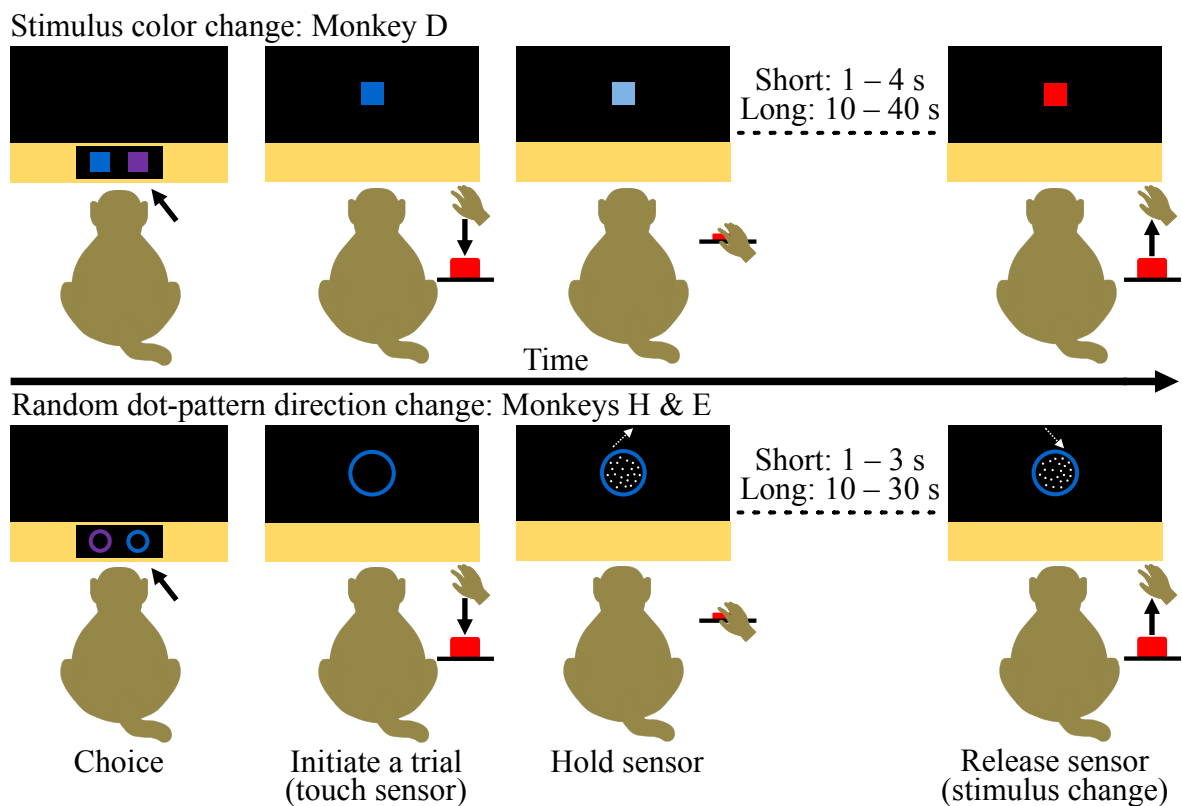


Figure 2.3: Time courses of the tasks provided during the Choice-based Severity Scale test. The monkeys indicated their choice by touching one of the task stimuli presented on the touchscreen of a cognitive testing system and were rewarded with 0.15 ml water to encourage engagement (first panel). The chosen stimulus appeared larger in the center of the touchscreen and blinked every 2.5 s until the monkey initiated a trial (second panel). Trials were initiated by the monkey touching and holding a proximity sensor (i.e., 'sensor; second panel). The stimulus either delimitated (stimulus color change task) or a random dot-pattern appeared moving in one direction (random dot-pattern direction change task) upon touch and the monkey had to hold the sensor until there was a second change in the stimulus (either another color change or change in the direction of the random dot-pattern; third panel). The duration of the hold depended on the monkeys' choice. Once the stimulus changed, the monkeys had to release the sensor within 2.5 s to receive the fluid reward associated with their choice (fourth panel).

The monkeys were given a choice on a trial-by-trial basis between the short and long hold task (Figure 2.3). Based on a brief pilot experiment that indicated biases due to stimulus position (see the section ‘Supplementary experiments’ in Appendix A), we set the position of the task stimuli to be deterministic so that the long hold task appeared in the opposite position as the last choice trial. The long hold task was rewarded with 15 ml more per correct trial than the short hold task (the two tasks also differed in the type of fluid reward; long: preferred juice; short: water). Training for the Choice-based Severity Scale test is described in the section ‘Preparation of the Choice-based Severity Scale test’ in Appendix A. Each monkey was tested for 10 days.

2.2.6 Statistical analyses

We analyzed our data using Generalized Linear Mixed Models (GLMMs) with a Bayesian framework via the ‘brms’ package (version 2.16.3: Bürkner, 2017) in R (version 4.1.2: R Core Team, 2021). brms calls Stan, a computational framework, to fit Bayesian models (Bürkner, 2017). Generally, we fit GLMMs with a binomial distribution and a logit-link function to investigate each monkeys’ training performance for the condition stimuli prior to the Choice-based Severity Assessment (model and results described in the section ‘Model description and results of the condition stimuli training for the Choice-based Severity Assessment’ in Appendix A) and their choice behavior during the CSS test.

For all GLMMs, fixed effects included in each model did not correlate above 0.5 (using Spearman’s correlation). We checked the distributions of model covariates and log transformed them when needed (i.e., trial number). Covariates were also z-transformed to a mean of 0 and a standard deviation of 1 to provide more comparable estimates and aid the interpretation of any interactions (Aiken et al., 1991; Schielzeth, 2010). We used weakly informative priors to improve convergence, avoid overfitting, and to regularize parameter estimates (McElreath, 2020). Binomial models had priors for each intercept that were a normal distribution with a mean of 0 and a standard deviation of 1. The priors for the beta coefficients were also a normal distribution with a mean of 0 and a standard deviation of 0.5. The priors for the standard deviation of group level (random) effects an exponential distribution with scale parameter 1. The priors for correlations between random slopes were LKJ Cholesky priors with scale parameter 2.

Each model was run using four MCMC chains for 2500 iterations, including 1000 “warm-up” iterations for each chain, with convergence of the chains confirmed by there being no divergent transitions, all Rhat values were equal to 1.00, and visual inspection of the plotted chains. We also checked model performance by using the ‘posterior predictive check’ (‘pp_check’) function from the ‘bayesplot’ package (Gabry & Mahr, 2022). We report model estimates as the mean of the posterior distribution with 95 % credible intervals (CI). To aid in the interpretation, we calculated the proportion of posterior samples that fell on the same side of 0 as the mean (Pr) to understand whether the fixed effects substantially influenced performance and choice behavior.

The Pr ranges from 0.5 to 1.0, where a Pr of 1.0 indicates a strong effect of a predictor (either negative or positive) and a Pr of 0.5 indicates no effect of a predictor on the response.

To investigate whether the monkeys developed a preference for one task over the other, we fit three GLMMs (one per monkey) with the response variable as whether the monkey chose the short or long hold task for each trial. We added session, the position of the monkeys' choice (left or right), and the amount of reward accumulated as fixed effects. We also included session as a random effect with all possible random slopes to allow the slopes to vary across sessions (Barr et al., 2013; Schielzeth & Forstmeier, 2009).

2.3 Results

2.3.1 Applying a Choice-based Severity Assessment

Through the Choice-based Severity Assessment protocol (see Choice-based Severity Assessment protocol) and experimental setup (Figure 2.2), we offered adult male rhesus macaques a choice between performing a basic experimental task in a cage or lab condition to generate a CSS. We found evidence of inter-individual differences in condition preference and how the monkeys responded to changes in the reward contingencies. During the first two phases, where the position of the cage condition was controlled for (changed from the upper quadrant to the lower quadrant, where the lab condition was positioned), monkey H exhibited a strong preference for the cage condition (100 % of six choice sessions each; Figure 2.4). Notably, this preference occurred despite the reward per trial being largely in favor of the lab condition (Figure 2.4). Once the reward per trial in the lower cage condition changed from grape juice to water (Figure 2.4), monkey H switched his preference to the lab condition (75 % of 18 choice sessions). Monkey H's preference for the lab condition persisted despite the reward per trial increasing to become largely in favor of the cage condition (Figure 2.4).

In contrast, monkey D exhibited an initial preference for the lab condition during the first bout of the first phase (upper cage vs lab condition, type of reward per trial was grape juice for both; Figure 2.4). Further into choice testing, however, monkey D switched his preference to the upper cage condition, irrespective of the amount of reward per trial in each condition (upper cage condition was chosen in 75 % of 16 choice sessions; Figure 2.4). Monkey D also chose the lower cage condition during the second and third phase (100 % of six choice sessions each), despite the location of the condition being controlled for, the type of reward per trial changing (cage: grape juice to water; lab: grape juice to banana juice), and amount of reward per trial being largely in favor of the lab condition (Figure 2.4). These data suggest that during the first phase monkey D sampled the different conditions, then settled on selecting the cage condition exclusively at the end of this first phase and continued to do so during the next two phases.

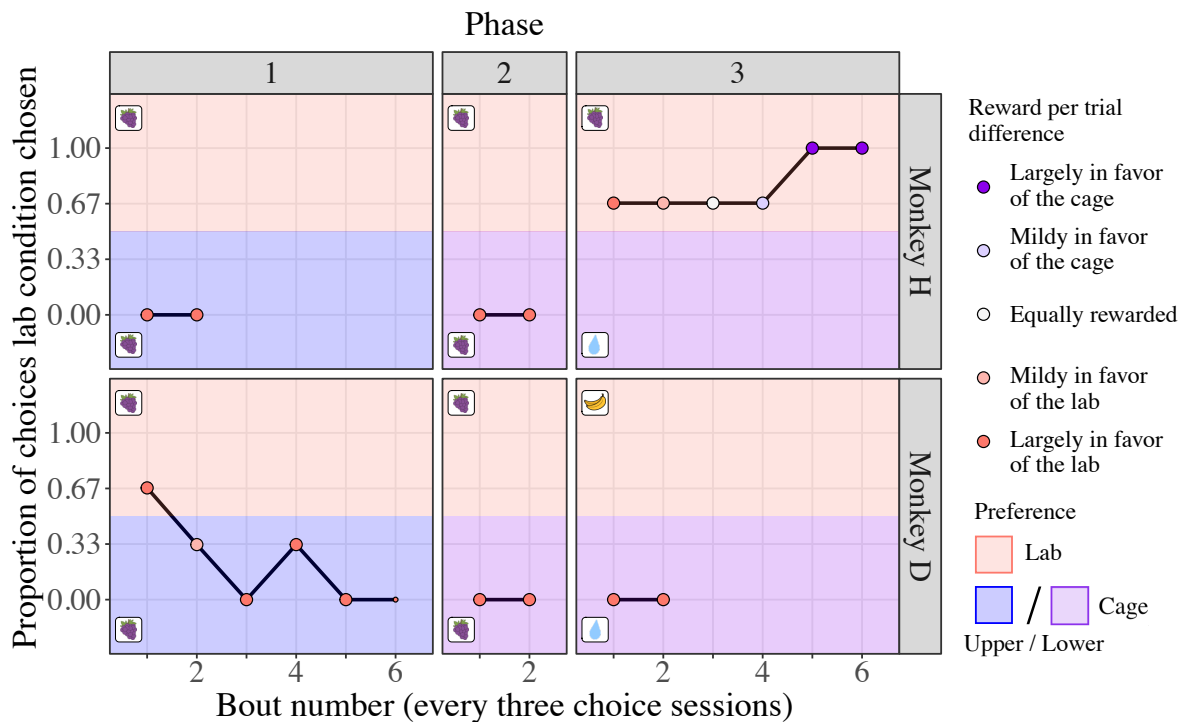


Figure 2.4: Results of the Choice-based Severity Assessment. The data are separated by each phase for the two monkeys that were tested in all phases. Two reference sessions (one per condition) preceded each choice session to remind the monkey of the consequences associated with each condition stimulus. One bout usually consisted of three consecutive choice sessions and their reference sessions, which took nine days to complete. Proportions were calculated for each bout (represented by point size, ranging from 1 to 3 choice sessions) to assess preference and adjust the reward per trial difference for the next bout accordingly. The reward per trial difference is indicated by point color. The type of reward used for each condition of each phase is indicated by the boxed picture on each panel. Over phases 1 and 2, grape juice (indicated by grapes) was delivered as a reward in each setting. During phase 3, the type of reward in the lab condition was changed to water (indicated by a drop of water) and the lower cage condition was changed to the monkey's preferred reward (monkey H: grape juice; monkey D: banana juice). A third monkey (monkey E) was only tested on the third phase and exhibited the same behavior as monkey D in phase 3 (see the section 'Results of the Choice-based Severity Assessment for the third monkey (E)' in Appendix A).

We tested an additional monkey (monkey E) during the third phase of the Choice-based Severity Assessment. Monkey E exhibited the same preference as monkey D, where he exclusively chose the lower cage condition (100 % of six choice sessions), despite the amount and type of reward per trial being largely in favor of the lab condition (see the section 'Results of the Choice-based Severity Assessment for the third monkey (E)' in Appendix A).

It should be noted that in our neuroscience setup the monkeys could easily compensate for lower reward per trial (typically experienced in the cage conditions) by performing more trials. Accordingly, the monkeys performed more trials on average in the cage conditions than the lab condition (lower cage: 666 ± 342 trials; upper cage: 964 ± 443 trials; lab: 97 ± 59 trials) when the fluid reward type was the same. Furthermore, the monkeys spent a greater amount of time working in the cage conditions than in the lab condition on average when the fluid reward type was the same (lower cage: 97 ± 33 minutes; upper cage: 94 ± 28 minutes; lab: 26 ± 9 minutes).

2.3.2 Choice-based Severity Scale test

The CSS test applied our CSS concept further by offering the monkeys choices between experimental tasks varying in trial duration. We found strong evidence that monkey H chose the long hold task more frequently overall, irrespective of the position of the task stimuli and session (Figure 2.5; Table 2.2). In contrast, there was strong evidence that the position of the task stimuli influenced the choice behavior of monkey D, where the long hold task was chosen less frequently when positioned on the left of the touchscreen (Figure 2.5; Table 2.2). Given that the position of the long hold task was deterministic due to our experience in previous pilot experiments (see the section ‘Supplementary experiments’ in Appendix A), our results suggest that monkey D had a left side bias, causing the short hold stimulus to appear on the left repeatedly. However, monkey D had to interrupt this bias to choose the long hold option, suggesting that these choices were deliberately made. Additionally, there was moderate evidence that session influenced the choice behavior of monkey D, where he selected the long hold task less frequently as the number of sessions increased (Figure 2.5; Table 2.2). There was little evidence that the choice behavior of monkey E was influenced by the position of the stimuli, but moderate evidence that he selected the long hold task more frequently as the number of sessions increased (Figure 2.5; Table 2.2). Such behavior suggests that, with additional sessions, monkey E learned that the tradeoff for engaging with the long hold task was more favorable with additional sessions.

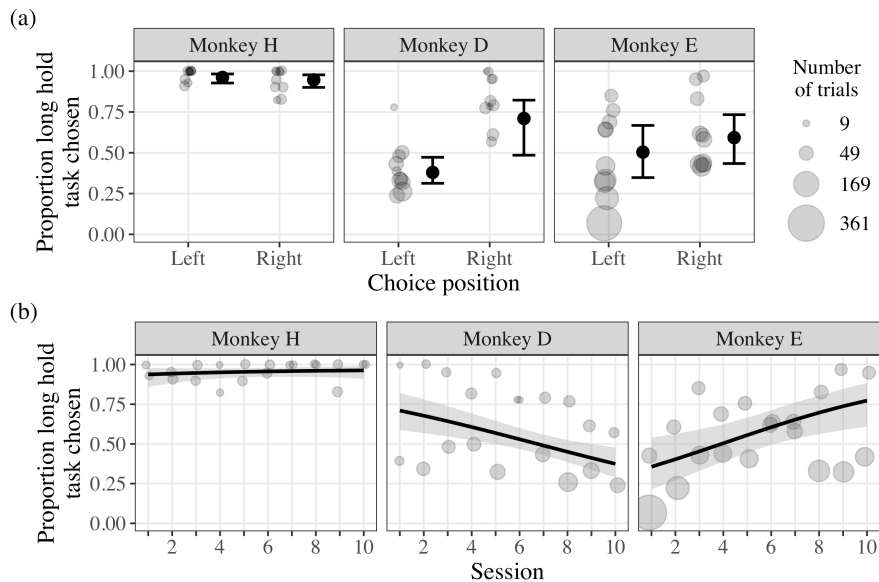


Figure 2.5: Results of the Choice-based Severity Scale test. The light grey points indicate the proportion of choices the long hold option was chosen over the total number for trials for each choice position, session, and monkey (range: 9 to 338 trials). In plot (a), the data are plotted by the position of the monkeys’ choices, where the large black points indicate the model probability estimates and the black whiskers indicate the 95 % credible intervals. In plot (b), the data are plotted by the proportion of trials the long hold task was chosen by session, where the black lines indicate the model probability estimates across sessions and the shaded grey areas indicate the 95 % credible intervals.

Table 2.2: Model results for the Choice-based Severity test. The binomial generalized linear mixed models for each monkey tested whether the short or long hold task was chosen.

Monkey		Estimate	SD	Lower CI	Upper CI	Pr
H	Intercept	3.29	0.38	2.55	4.05	1.00
	Choice position (right) ^a	-0.34	0.39	-1.09	0.42	0.81
	Session	0.22	0.28	-0.32	0.80	0.78
	Trial number	-0.21	0.25	-0.70	0.27	0.80
D	Intercept	-0.49	0.17	-0.79	-0.11	0.99
	Choice position (right) ^a	1.42	0.42	0.43	2.04	1.00
	Session	-0.31	0.14	-0.59	-0.03	0.98
	Trial number	-0.05	0.14	-0.34	0.20	0.64
E	Intercept	0.02	0.33	-0.63	0.70	0.51
	Choice position (right) ^a	0.37	0.39	-0.43	1.09	0.83
	Session	0.59	0.27	-0.03	1.05	0.97
	Trial number	0.18	0.16	-0.15	0.48	0.88

Estimate: slope of the predictor. SD: standard deviation of the estimate. CI: 95 % credible interval. Pr: proportion of the posterior samples that fall on the same side of 0 as the mean.

^aLeft was the reference level for choice position.

Monkey H was the most efficient (i.e., least effort for most reward) and received 14.4 ml per trial on average across sessions, whereas monkey D and monkey E received 7.8 and 7.1 ml per trial respectively on average across sessions. While these descriptive statistics suggest that the strategies monkey D and E employed were not optimal, the monkeys were still able to receive over 400 ml reward per session on average by choosing the long hold option occasionally and engaging in more trials. Thus, it may have not been necessary for the monkeys to exclusively choose the long hold option throughout the session.

2.4 Discussion

Objective assessment of lab animal welfare is crucial not only for ensuring that high standards of animal welfare are maintained, but for the validity and quality of the scientific experiments they are involved in (Jennings & Prescott, 2009; Poole, 1997). To address the issues of ranking and scoring animal welfare parameters, we proposed and tested our Choice-based Severity Scale

(CSS) concept by giving adult male rhesus macaques a series of choices to perform a basic experimental task close to their home cage (cage condition) or in a laboratory environment (lab condition). The data we collected consistently support the validity of the CSS concept, where we find distinct preferences for the conditions that we provided the monkeys and that these preferences can be influenced by changes in reward contingencies. During the Choice-based Severity Assessment, we limit the potential influence from the experimenter and/or environment on the monkeys' choices by providing the monkeys choices between visual stimuli associated with the conditions (i.e., condition stimuli) on a neutral cognitive testing system. We provide guidelines in Appendix A (see the section 'Guidelines for Choice-based Severity Assessments in animals') to highlight several points (e.g., training, experimental setup, prior experience) to consider during the design of Choice-based Severity Assessments. Collectively, we believe that our study provides a basis for expanding and adapting the CSS concept to other species and other conditions than those we have explored in this study.

A core tenant of a Choice-based Severity Assessment is that it can be applied individually. In support, the individual monkeys' choice behavior during our Choice-based Severity Assessment indicates that the CSS is indeed sensitive to inter-individual differences. During our Choice-based Severity Assessment, one monkey switched to the less preferable condition (i.e., lab condition) given a large enough reward difference (juice instead water) in favor of that condition. Interestingly, the same monkey responded the strongest to the difference in reward amount per trial during the test of the CSS. While this behavior contrasts that of the others, it highlights that these monkeys may have different point of subjective equality for the costs and benefits associated with the conditions we tested (upper cage vs. lab, lower cage vs. lab). In other words, there may have not been enough incentive for the other two monkeys to select the less desirable condition due to the flexible time window to work in each setting and/or they may have not noticed the changes in reward contingencies (discussed in more detail later). It is well known that individuals respond differently to their internal and external environments as aspects of their life histories differ (e.g., species, age, sex, personality: Coleman, 2012; Izzo et al., 2011; Palmer et al., 2022; Sloan Wilson et al., 1994). Such differences are important to consider when designing the ranking and scaling of welfare parameters as animals do not perceive and experience welfare conditions in the same way. The CSS represents a severity assessment tool that matches this requirement.

There are several explanations for why the amount of reward did not influence choice behavior during the Choice-based Severity Assessment. Given the regulatory requirement to provide ample time to collect as much reward as desired, the additional reward per trial might not have been enough incentive to choose the lab condition. Even though the lab condition is the most efficient way to gain fluid reward and return to the home enclosure earlier, the monkeys could easily compensate this by performing more trials in the cage condition. Alternatively, detection of reward contingency changes may have been hindered by the 10-minute delay between the selection of a condition stimulus and its corresponding consequences. This 10-minute delay was necessary to transport the monkeys to the location of the lab condition and was matched

with a waiting period 10-minute in the time course of the cage condition. Within these delay periods, multiple distracting events could occur (e.g., transport to lab, social group interactions) that may have made the formation of an association between each stimulus and its outcome more challenging.

Our CSS testing data show that all three monkeys engage in the long-hold task when the reward per trial was substantially higher than the short-hold task. These data support the core approach of the CSS that preference between two conditions can be reversed using reward amount. Thus, reward amount can be used a common unit to scale conditions across different parameters and domains in a comparable and objective way. Given that we were able to reverse preference using reward amount for the CSS test and not the Choice-based Severity Assessment, differences in reward amount may be easier for animals to detect when the delay between a choice and its consequences is short (e.g., delay was 40 seconds for the CSS test vs. 10 minutes for Choice-based Severity Assessment).

We recognize that giving the monkeys choices between the complex, full-scale experimental conditions in our study was time intensive. But as the CSS protocol closely reflected the actual procedures of the lab and cage conditions, we could build an accurate picture of how these conditions were experienced by the monkeys due to their choice behavior in the Choice-based Severity Assessment. Other conditions may not be as time intensive to determine animal preferences because visual stimuli may not be needed to represent each condition, which necessitate training sessions to remind the animal of the consequences of each condition stimulus. For example, offering choices between different types of bedding or enrichment devices would not require the items to be associated with species relevant stimuli because the items themselves could be offered simultaneously.

Given the benefit of its animal-centric approach, the CSS concept should be validated and developed further as a powerful animal welfare assessment. Naturally, testing more individuals is a good first step forward. Further validation by other individual-based welfare parameters such as physiology (e.g., heart rate variability: von Borell et al., 2007), stress hormones (e.g., Pfeifferle et al., 2018), blood values (e.g., Wegener et al., 2021), and behavior (e.g., abnormal: Gottlieb et al., 2013a) is also warranted. Another interesting comparison would be to offer a condition that is putatively more positive in valence. In our laboratory, performing the basic experimental task in the home cage itself, where the monkeys have full visual access to conspecifics and can engage in other behaviors like foraging, is an alternative, putatively more positive, condition that could be compared. Conditions should be associated with species-relevant stimuli and a CSS protocol can be created to accommodate such conditions (Kahnau et al., 2020). For example, different compartments can be associated with different conditions (e.g., conditioned place preference tests comparing, e.g., food and an aversive procedure: Millot et al., 2014; social partners: Panksepp & Lahvis, 2007; analgesic drugs: Roughan et al., 2014) and offered simultaneously to animals to determine preferences. Lastly, expanding the CSS concept to test other species and other conditions warrants exploration.

Summary

Historically, animal welfare science has shied away from recognizing animals' subjective experiences as meaningful to their welfare but interest in linking the two topics has grown in the last few decades (Marchant-Forde, 2015). The CSS concept that we propose here has fundamental benefits for making welfare and severity assessments less anthropocentric and more animal-centric by shifting the perspective of lab animals into the central focus. To our knowledge, our study is the first to offer lab animals choices between experimental procedures. In summary, the CSS is a powerful tool that can help shape the refinement of husbandry and research practices (Schapiro & Lambeth, 2007), and thus strengthen the validity and quality of scientific research.

Chapter 3

Adult male rhesus macaques (*Macaca mulatta*) can associate abstract stimuli with long-delayed reinforcement

Lauren C. Cassidy^{a,b}, Alexander Gail^{a,b,c}, Stefan Treue^{a,b,c}, and Dana Pfefferle^{a,b}

^aWelfare and Cognition Group, Cognitive Neuroscience Laboratory, German Primate Center - Leibniz Institute for Primate Research, Göttingen, Germany; ^bLeibniz-ScienceCampus Primate Cognition, Göttingen, Germany; ^cFaculty for Biology and Psychology, University of Göttingen, Göttingen, Germany

The following manuscript is in preparation for publication.

Contribution to the field

The process of learning to associate a choice to its consequences occurring substantially later is cognitively demanding. While long-delay learning with positive reinforcement in animals is possible using interoceptive stimuli and navigation signals, it is less clear if they can do so for abstract stimuli. Our study suggests that monkeys are capable of long-delay learning with such stimuli and delayed positive reinforcement up to 10 minutes, even when the stimuli are *novel*. Our paradigm included a dynamic stimulus that provided information about the choice made and the passage of time. Such information would be helpful for the animal to determine whether the reward is worth waiting for but is typically lacking from tasks testing delay tolerance. We suggest that tasks assessing preference and continued task engagement validate long-delay learning capabilities better than those lacking these features. The following study also provides support for the interpretations in Chapter 2, where a 10 minute delay between the monkeys' choice and its consequences occurred.

Author contributions

All authors contributed to the conception and design of the study. Lauren Cassidy programmed the task. Alexander Gail and Stefan Treue provided the monkeys, test systems, and infrastructure support. Lauren Cassidy trained the monkeys and collected the data. Lauren Cassidy performed the statistical analyses. All authors interpreted the data. Lauren Cassidy wrote the manuscript, with revision by Dana Pfefferle, Alexander Gail, and Stefan Treue.

Acknowledgments

We thank the German Primate Center animal care and veterinary staff for taking care of the monkeys. We also thank Ralf R. Brockhausen for task programming advice.

Statement of ethics

Research with non-human primates represents a small but indispensable component of neuroscience research. The scientists in this study are aware and are committed to the great responsibility they have in ensuring the best possible science with the least possible harm to the animals (Roelfsema & Treue, 2014; Treue & Lemon, 2022).

All animal procedures were conducted according to the relevant national and international laws and guidelines, including the German Animal Protection Law, the European Union Directive 2010/63/EU on the Protection of Animals used for Scientific Purposes and the Society for Neuroscience Policies on the Use of Animals and Humans in Neuroscience Research, and were approved by the responsible regional government office (Niedersaechsisches Landesamt fuer Verbraucherschutz und Lebensmittelsicherheit, LAVES) under the permit numbers 33.19-42502-04-18/2823.

Conflict of interest statement

This research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding sources

This study was supported by the German Research Foundation Research unit 2591 “Severity assessment in animal-based research” assigned to ST, AG, and DP (grant numbers: TR 447/5-1/2, GA 1475/6-1/2, PF 659/5-2). We acknowledge support by the Leibniz Association through funding for the Leibniz ScienceCampus Primate Cognition.

Supplementary material and data

See Appendix B for the associated supplementary material. The data supporting this article will become available upon publication. Code for the delayed reinforcement task can be provided on request.

Abstract

Associating a choice to its consequences occurring substantially later is cognitively demanding. Information needs to be maintained in memory across the delay and there is an increased likelihood of intervening events, making the association less apparent. The nature of a stimulus influences if and how quickly associations with delayed positive reinforcement can be learned. Animals can learn long-delay associations using interoceptive stimuli and navigation signals, but it is less clear if such learning is possible with abstract stimuli. Therefore, we designed a task offering adult male rhesus macaques (*Macaca mulatta*) a choice between two abstract stimuli, associated with low or high fluid reward that was delivered up to 10 minutes after their choice. In the first experiment, we increased delay in a stepwise fashion to 10 minutes for the same set of two stimuli. In a second experiment, we set delay to discrete durations (2, 6, or 10 minutes) for *novel* stimulus sets. We found that the monkeys developed a preference for the high reward stimulus over the low reward stimulus, despite having to wait for up to 10 minutes in the first experiment. This preference was independent of incremental training as the same preference pattern was found when stimuli were novel, and the delay was fixed in the second experiment. Furthermore, the monkeys were more likely to complete trials where they had selected the high reward stimulus and to abort low reward trials. This selective behavior suggests that the monkeys sustained their commitment to a high reward decision and retained information about the quality of the chosen abstract stimulus over the course of the delay. Our findings suggest that rhesus macaques can use abstract stimuli to make informed value-based choices even if the consequence of their choice is delayed by several minutes. We also discuss leveraging these long-delay learning capabilities to offer captive animals' choices between options involving a component of delay such as husbandry or experimental procedures.

Keywords: delayed reinforcement, associative learning, long-delay, preference, choice

3.1 Introduction

Animals take in a plethora of sensory information while foraging for food. This external and internal sensory information guides feeding decisions to be more effective. However, the usefulness of a particular stimulus is affected by time and space (reviewed for non-human primate species in Dominy et al., 2001), not to mention the sensory adaptations of the animal itself (e.g., nocturnal species may rely more heavily on olfaction than diurnal species: Bicca-Marques & Garber, 2004). Linking nearby resources with properties such as shape or color, can be easily conditioned as the animal receives feedback about the quality of the resource more or less immediately. Yet resources are often far away in time and space, so animals must make decisions based on the information they currently have. In such situations, foraging would be facilitated if animals can learn associations between abstract stimuli (e.g., landmarks) and temporally and/or spatially distant resources. Presently, it is unclear if animals are capable of learning such long-delay associations.

Delay can have a substantial, usually detrimental, impact on decision-making and learning in animals (reviewed in Lattal, 2010). Studies of associative learning indicate faster learning and stronger associations when the delay between an action (behavior) and its resulting consequence is short – on the order of seconds – marked by a distinct stimulus change (i.e., signal or secondary reinforcer, e.g., Azzi et al., 1964; Lieberman et al., 1979; Richards, 1981; Thomas et al., 1983). Extending the delay period to durations of minutes and hours (i.e., long delays), however, increases the likelihood that other more complex and independent events may occur (intrinsic and/or external) and impede learning (Revusky, 1971). Ample research in animals has demonstrated long-delay learning capabilities of stimuli that result in potentially life-threatening events, where adaptation is an evolutionary necessity (e.g., visual properties of poisonous fruit, appearance of poachers or hunters, reviewed in Bernstein, 1999; LeDoux, 2003). Contrary to such avoidance learning of strongly aversive stimuli, learning from choices involving delayed positive outcomes (e.g., food provision) represents a situation where less direct selective pressure is involved. Despite less selective pressure, an animal can still reap a benefit to a greater or lesser extent and satisfy immediate needs (e.g., hunger) by associating stimuli with certain outcomes. The benefits of optimally choosing the higher value outcome would become apparent if the difference in value is substantial and/or once the beneficial effects have accumulated over the long-term. Such a circumstance begs the question, under what conditions is it possible for an animal to form associations between stimuli and their long-delayed positive outcomes?

Stimulus modality likely plays an important role in establishing a link to delayed positive consequences as animals use their different sensory adaptations to navigate their environment and perceive desired resources. Interoceptive stimuli, such as flavor, have been successfully linked to delayed positive reinforcement across several different species of animals. For example, rats (*Rattus species*) and sheep (*Ovis aries*) developed preferences for flavors associated with more nutritive diets despite the nutritive effects occurring 10 minutes or more

later (e.g., Arsenos et al., 2000; Baker & Booth, 1989; Capaldi et al., 1987; Pérez et al., 1995). Navigation signals (i.e., spatial orientation) also play an important role in learning as the memory of previous food sources can influence decision-making to maximize foraging efficiency by reducing the time and energy invested (reviewed by Garber, 2000). Consequently, conditioned place preference tests using T-mazes in capuchin monkeys (*Cebus apella*) and rats have demonstrated that baseline branch preferences could be reversed in one trial even though delayed reinforcement was delivered 30 minutes later (D'Amato & Puopolo, 1981; Safarjan & D'Amato, 1981). While several sensory modalities may be useful for foraging, the relative salience of exteroceptive stimuli (auditory, olfactory, visual, tactile) for the species being tested may influence how quickly and if associations can be learned. Distinguishing stimuli with delayed positive outcomes by visual characteristics, for example, may better facilitate long-delay learning for animals with highly developed visual systems such as diurnal non-human primates. For these species, discriminating stimuli by their abstract visual properties is adaptive as components such as brightness or color can enhance the selection of preferred or avoidance of unpleasant foods (Hernández et al., 2021; Riba-Hernández et al., 2005; Sánchez-Solano et al., 2020).

Little scientific literature exists explicitly testing the development of associations between abstract stimuli and their delayed positive consequences in non-human primates. Ferster & Hammer (1965) found that rhesus macaques (*Macaca mulatta*) and baboons (*Papio papio*) learned to respond to colored keys associated with a delayed reward, as signaled by a blackout, by manipulating the amount of food delivered and number of key presses needed to receive reward. While engagement in the task reduced substantially as the delay increased, the results demonstrate evidence of learning when the delay was extended up to 24 hours and that gradually increasing the delay was not a precondition for learning. However, in these experiments the monkeys were not offered a choice between different options. Testing a preference between different options would not only reveal how valuable certain resources are in relation to one another (e.g., Hosey et al., 1999), but additionally demonstrate an understanding of the association between an abstract stimulus and its delayed outcome when preferences are apparent. In a task offering a choice, D'Amato et al. (1981) demonstrated three of four capuchin monkeys were not able to acquire a preference in a T-maze task involving a 30-minute delay via visual discrimination. Preferences in the T-maze task may not have developed due to pre-existing side biases that may have overshadowed the information from the added visual stimuli in the experiment, thus interfering with long-delay learning. Collectively, these studies indicate mixed evidence for long-delay learning with abstract stimuli in non-human primates and further exploration of these capabilities is warranted. Such capabilities could be tested for more rigorously by offering non-human primates multiple options to choose from, where abstract stimulus information is not eclipsed by other biases.

We were interested in testing if adult male rhesus macaques could develop preferences for abstract visual stimuli and their positive consequences (i.e., fluid reward), despite an interim delay of up to 10 minutes. Our task gave the monkeys a choice between two abstract stimuli that

were associated with a low or high value reward that was delivered following a predetermined delay. Like other long-delay learning experiments, we considered learning to occur if the monkeys developed a preference for one stimulus over the other and continued to successfully reach the end of delay period (i.e., were present to trigger the associated reward). As an alternative strategy, the monkeys could abort trials frequently to receive a small reward only by that delivered for making a choice. Opting for such a strategy would be similar to findings on studies of patience and self-control, where non-human primates generally opt for receiving small rewards sooner rather than larger ones later (e.g., Evans & Beran, 2007; Rosati et al., 2007; Stevens et al., 2005).

In our first experiment (*fixed stimuli experiment*), we increased delay in a stepwise fashion dependent on the monkeys' reaching the consequences of their choices (independent of which stimulus was selected). Generally, we expected that the monkeys would exhibit a preference for the high reward stimulus when delay was relatively short (under 1 minute). However, we did not predict that the monkeys would reach the maximal delay (10 minutes), but rather opt for an alternative strategy (e.g., abort trials frequently). In our second experiment (*generalized stimuli experiment*), we introduced sets of novel stimuli pairs for three different delay periods (2, 6, and 10 minutes) to determine if effects were independent of incremental training that may have occurred in the first experiment. If the monkeys were able to wait 10 minutes to receive the consequences of their choice in our (first) fixed stimuli experiment, we expected that they could apply this ability to novel stimuli and exhibit a preference for the high reward stimulus. Choice and trial outcome (whether the monkeys completed, uncompleted, or aborted a trial) behavior were investigated with regards to delay across both experiments.

3.2 Materials and methods

3.2.1 Study subjects and housing facility

We conducted the study on two adult male rhesus macaques (9 and 10 years old at the start of the experiment) housed in isosexual pairs with visual and auditory contact to other macaque groups at the German Primate Center, Goettingen, Germany. Both monkeys had previous training for other cognitive tasks using a touchscreen. Housing compartments were carpeted with wood shavings and were furnished with fixed and dynamic perching (e.g., platforms, chains) and environmental enrichment Cassidy et al. (2021). Each housing space was comprised of two large rooms, substantially exceeding the size requirements set by EU directive 2010/63/EU. One was an indoor room with a 12-hour light/dark cycle (from 07:00 to 19:00) connected by an elevated tunnel to an outdoor-exposed room, where monkeys could experience natural weather fluctuations. Monkeys had monkey chow, fresh fruits and vegetables, and water *ad libitum* on days where they were not being tested. On testing days, fluid consumption was based on the monkeys' performance in the behavioral tasks (see Figure 3.1, fluid control practices are

described in detail in: Pfefferle et al., 2018). We monitored the monkeys' weight everyday of testing. Additionally, monkeys were monitored daily by veterinarians, monkey facility staff, and the lab's researchers who all have specialized training for working with non-human primates. See 'Statement of ethics' for permit information.

3.2.2 Cognitive task

3.2.2.1 Experimental testing apparatus and software

For the behavioral tasks we used a touchscreen system (eXperimental Behavioral Instrument, XBI) developed within the lab for cage-side cognitive task training and testing (Berger et al., 2018; Calapai et al., 2017). In addition to running complex tasks, the XBI can deliver two types of fluid reward (water and grape juice in our experiments) to non-human primates via a reward tube positioned in front of the touchscreen (30.4 cm by 22.7 cm). The position of the reward tube encourages the monkeys to adopt stereotypical postures when engaging with the device (Calapai et al., 2017). We programmed the tasks in MWorks (version 0.9; <http://mworksproject.org/>), an open-source C++ -based software for real-time controlled behavioral tasks (Calapai et al., 2017). The XBI was mounted to the 'bay area' (extension from the main caging) of the monkeys' indoor room, allowing the monkeys' to freely move around and engage in other activities.

3.2.2.2 Delayed reinforcement task

We tested the monkeys in two experiments using the same general task structure (i.e., delayed reinforcement task) as depicted and described in detail in Figure 3.1. Trials were initiated by the monkeys by touching a colored square (3.7 cm by 3.7 cm, '*start button*') located within a colored bar (30.4 cm by 7.6 cm) present on the lower third of the touchscreen. Once the start button was touched, two stimuli (each fitting within a 5.2 cm circle) differing in shape and color appeared counterbalanced on the left and right side, respectively, of the screen. The monkey could select either stimulus by touching it. Trials where no selection was made were coded as "no decision" trials. Each stimulus was associated with a different type and amount of reward (i.e., low reward, high reward); the amount of the high reward stimulus was dependent on how long the monkey needed to wait to receive the consequences of his choice (see Table 3.1). Once the monkey selected a stimulus, both stimuli disappeared and the monkey was rewarded with 0.15 ml of water (i.e., choice reward). This choice reward was given to encourage stimulus selection. In addition, this choice reward provided the monkey with the possibility to abort trials (i.e., not waiting for the trials consequences). Therefore, the monkeys could abort trials repetitively to receive fluid rather than having to wait for the consequences of the stimulus. If the monkey made a choice, an "expanding clock" appeared in the center of the black background in the shape of the chosen stimulus. The expanding clock consisted of the shape of the chosen stimulus reduced in size (fitting within a 0.5 cm diameter circle) and transparency nested within

a larger, grey version of the shape (fitting within a 12.2 cm diameter circle). With advancing time, the size of the nested shape increased linearly with the length of the delay until it reached the same size as the grey shape. The length of the delay was dependent on the experiment being conducted (see next section and Table 3.1). Therefore, the shape increased faster if the delay was short (e.g., 1 s) than if the delay was longer (e.g., 1 min), offering the monkeys a means of determining how long they must wait for the consequences of their choice. Monkeys had the option of waiting for the corresponding consequences of their choice by touching the expanding clock once it had completely filled and was illuminated (i.e., completed trial; response window: 2 s), aborting the trial by touching the expanding clock (i.e., aborted trial) before it was filled and illuminated, or missing the 2 s window during which the illuminated expanding clock could be touched for reward (i.e., uncompleted trial).

3.2.2.3 Task training

We trained the monkeys on the contingencies of the expanding clock prior to starting the experiments. The central contingency of the expanding clock was that the monkeys had to wait for the internal shape stimulus to expand and illuminate before touching it to receive the consequences (fluid reward according to waiting time). During the first training step, the monkeys were presented with one stimulus at a time, which they needed to touch to view the expanding clock with the corresponding shape and color. Simultaneously, we incrementally increased the transparency of the expanding clock. Over the course of training, we increased the delay of the consequences from 200 ms to 1000 ms in increments of 20 ms. Changes in delay and transparency occurred when the monkeys had completed three trials in a row until a delay of 1000 ms was reached. After this step, the task changed to the state tested in the fixed stimuli experiment (see Figure 3.1) to familiarize the monkeys with making choices.

3.2.2.4 Fixed stimuli experiment

In the fixed stimuli experiment, we increased the delay from 0.2 to 10 min in a stepwise fashion after the monkey had successfully completed three trials in a row of any choice (i.e., experimental staircase; see Table 3.1). Following this criterion, increases in delay were 100 ms at minimum, then by increments of 10 % of the last trial's delay, up to a maximum increment of 1000 ms. Once a delay of 7 min of was reached within the task of the fixed stimuli experiment, we increased the delay step size to 1 min. The shape and color of the low and high reward stimuli stayed the same throughout this experiment.

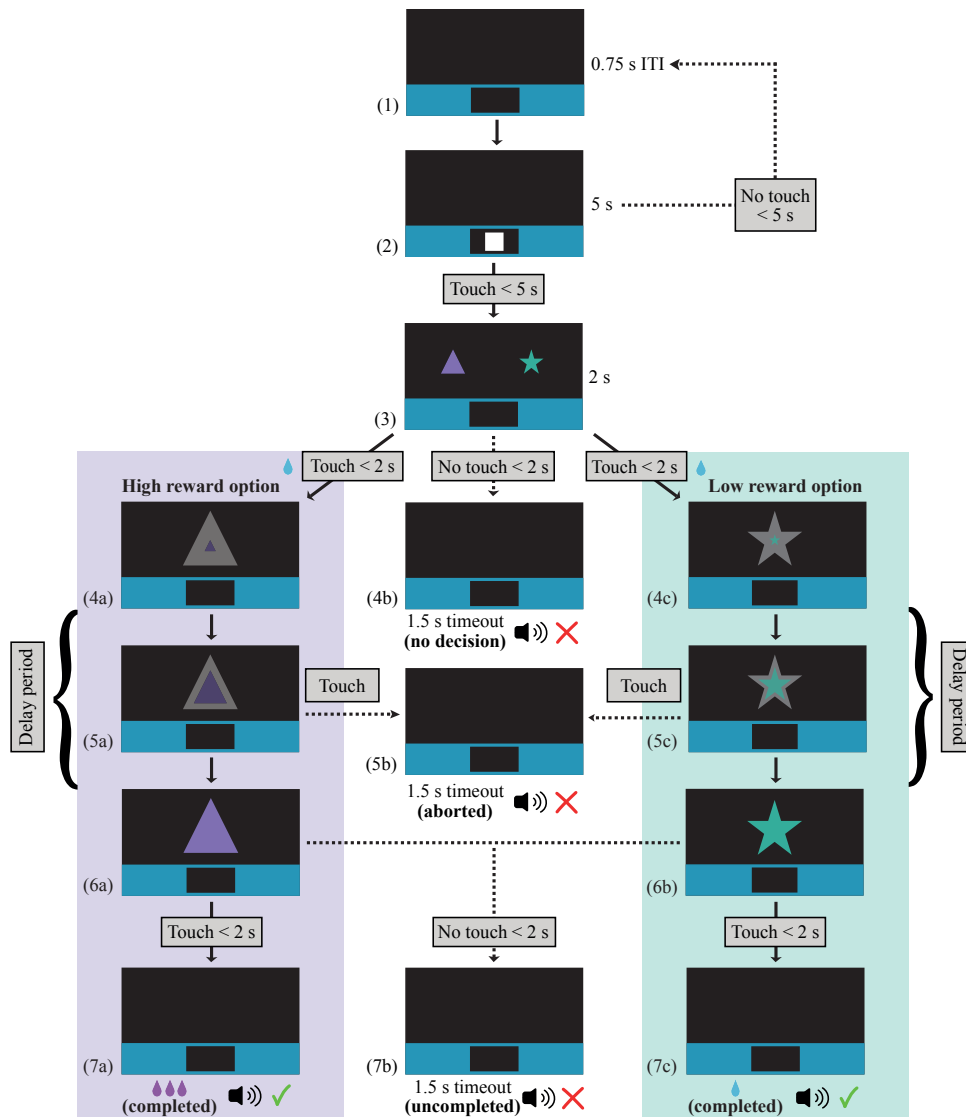


Figure 3.1: General time course of the experiment. (1) Each trial was separated by a 0.75 s inter-trial interval (ITI), where no stimuli were presented on the screen. (2) A white square (i.e., ‘start button’) was presented for 5 s or until the monkey touched. (3) Two stimuli were presented (position counterbalanced across trials) for 2 s or until a stimulus was touched. (4a/4c) Once a stimulus was touched, monkeys were rewarded with drop of water and a small, delimited version of the shape appeared within a larger grey version. (4b) If no stimulus was touched, both disappeared, an error sound (buzz) was heard, followed by a 1.5 s timeout (i.e., no decision trial). (5a/5c) The size of the delimited stimulus expanded at a speed linear to the scheduled delay (i.e., delay period). The delay period during the fixed stimuli experiment increased continuously from 0.02 to 10 min, whereas delay was set to the discrete values of 2, 6, and 10 min during the generalized stimuli experiment. (5b) If the expanding stimulus was touched during the trial, an error sound occurred, followed by a 1.5 s timeout (i.e., aborted trial). (6a/6b) The central stimulus illuminated once it reached its full size, indicating that the delay period had finished. (7a/7c) If the monkey touched the illuminated stimulus within 2 s, then he received the corresponding reward (water for low reward stimulus, grape juice for high reward stimulus) and a correct (ding) sound was heard (i.e., completed trial). (7b) If the monkey did not touch the illuminated stimulus within 2 s, then an error sound occurred, followed by a 1.5 s timeout (i.e., uncompleted trial).

3.2.2.5 Generalized stimuli experiment

Following the fixed stimuli experiment, we designed the generalized stimuli experiment, including *novel* stimuli, to determine if effects were independent of incremental training. We set the delay to 2, 6, or 10 min for the generalized stimuli experiment, which were tested in that order. No training condition preceded the testing of each delay in the generalized stimuli experiment. Each delay had four different sets of stimulus pairs (i.e., stimulus sets) that were tested over several days (12 stimulus sets in total; see Table 3.1). As neither monkey had seen the stimuli of the generalized stimuli experiment before, they first had to experience the stimulus choice consequences to figure out which was associated with low and high reward. Therefore, we coded each trial depending on the information state of the monkey (see Table 3.2). We considered the monkeys to be uninformed of the consequences of choosing a stimulus from a stimulus set until they had successfully completed at least one choice of each stimulus (i.e., informed).

Table 3.1: Experimental testing information.

Exp.	Delay [min]	Sessions tested [day]	Testing duration [h]	Stimulus sets	Low reward [ml wat]	High reward [ml grj]
Fixed	0.2 - 10	16	2	–	0.15	0.6 + 0.2 per s delay
General.	2	1-2 per stimulus set	2	4	0.15	12.2
	6	2 per stimulus set	3	4	0.15	36.2
	10	4-5 per stimulus set	3	4	0.15	60.2

The delay in the fixed stimuli experiment was increased only after 3 trials were completed in a row. Following this criterion, delay was increased at a minimum increment of 100 ms, then by increments of 10 % of the last trial's delay, up to a maximum increment of 1000 ms. Once a delay of 7 min of was reached, delay was increased by increments of 1 min. General.: generalized. wat: water. grj: grape juice.

3.2.3 Experimental testing protocol

The data for this study were collected between June and September 2020. Monkeys were tested on weekdays for two hours, beginning between 14:15 and 15:15. Prior to each testing session, we weighed each monkey. For testing, we temporarily separated the monkey's housing partner to ensure the correct identity of the individual engaging with the XBI. Monkeys still had visual

and auditory contact with other macaques during the testing sessions and were free to move around and engage in other activities. We tested both monkeys on the fixed stimuli experiment (increasing delay) first and then on the generalized stimuli experiment (pre-determined delays with four novel stimulus sets each). We did not counterbalance the order of the experiments as we did not know if it was possible for the monkeys to complete the first experiment initially.

3.2.4 Statistical analyses

We conducted two types of analyses on choice and trial outcome behavior using a Bayesian framework for each experiment conducted in our study, resulting in four Bayesian generalized linear mixed models (GLMMs). Choice behavior (choice models 1 and 2), whether the monkeys' chose the high or low reward stimulus, and trial outcome behavior (trial outcome models 1 and 2), whether the monkeys' completed, uncompleted, or aborted a trial, were our response variables (Table 3.2). These analyses were used to answer the following questions:

(Q1) Do the monkeys learn to wait up to 10 min for the consequence of a choice between abstract stimuli (fixed stimuli experiment)?

(Q2a) Is the monkeys' choice behavior modulated by increasing delay (fixed stimuli experiment; choice model 1)?

(Q2b) Is the monkeys' choice behavior modulated by increasing delay, despite stimulus pairs being *novel* (generalized stimuli experiment; choice model 2)?

(Q3a) Is the monkeys' trial outcome behavior influenced by their choice or/and delay (fixed stimuli experiment; trial outcome model 1)?

(Q3b) Is the monkeys' trial outcome behavior influenced by their choice or/and delay, despite stimulus pairs being *novel* (generalized stimuli experiment; trial outcome model 2)?

We analyzed our data using the 'brms' package (version 2.16.3: Bürkner, 2017) in R (version 4.1.2: R Core Team, 2021). brms calls on Stan, a computational framework, to fit Bayesian models (Bürkner, 2017). Each model was run using four MCMC chains for 2500 iterations, including 1000 "warm-up" iterations for each chain. We checked convergence diagnostics for each model, finding there were no divergent transitions, that all Rhat values were equal to 1.00, and that visual inspection of the plotted chains confirmed convergence. We used weakly informative priors to improve convergence, avoid overfitting, and to regularize parameter estimates (McElreath, 2020). The prior for each intercept was a normal distribution with a mean of zero and a standard deviation of 1. For the beta coefficients, we used a prior with a normal distribution with a mean of 0 and a standard deviation of 0.5. For the standard deviation

of group level (random) effects, we used a prior with an exponential distribution with scale parameter 1. Lastly, we used a LKJ Cholesky prior with scale parameter 2 for the correlations between random slopes.

To test if the monkeys preferred the high reward over the low reward stimulus (Q2a, Q2b), we fit a GLMM for each experiment with a logit-link function and the family specified as ‘binomial’, as choice behavior was a binary outcome (high or low reward stimulus) where each choice was one trial. Delay was our main test predictor for both experiments (fixed stimuli experiment: 0.2 to 10 min; generalized stimuli experiment: 2, 6, and 10 min) to test if choice behavior was modulated by how long the monkeys had to wait to receive the consequences of their choice. Monkey identity (D, H), choice position (left, right), reward accumulated, and trial number were added to these models as control predictors as these variables may have influenced choice behavior (Table 3.2). We checked all variables included in each model for correlations (Pearson correlation coefficients below 0.5).

Additionally, we checked the distributions of model covariates and log transformed trial number to be normally distributed. Then we z-transformed covariates to a mean of 0 and a standard deviation of 1 to provide more comparable estimates and aid the interpretation of any interactions (Aiken et al., 1991; Schielzeth, 2010). We included session identity as a random effect for the model of the fixed stimuli experiment with all possible random slopes to keep type I error rates at the nominal level of 0.05 (Barr et al., 2013; Schielzeth & Forstmeier, 2009). As a random effect for the second model, we included session nested within stimulus set as we tested multiple days for each stimulus set in the generalized stimuli experiment; all possible random slopes were included.

As the two monkeys may have responded differently to the task, we considered an interaction between monkey identity and delay using leave-one-out (LOO) cross-validation by applying Pareto-smoothed importance sampling (PSIS, Vehtari et al., 2017). Specifically, we compared and checked models fit with and without the interaction between delay and monkey identity for each choice behavior model using the `loo` package (version 2.4.1: Vehtari et al., 2020). Each model was checked for PSIS estimates over 0.7 as these may have allowed the predictive performance model to be overestimated (McElreath, 2020; Vehtari et al., 2017; Vehtari et al., 2020). Models with PSIS estimates larger than 0.7 were refit by leaving out problematic observations one at a time and recalculating the LOO approximation (Vehtari et al., 2017). We considered models when their expected log predicted density (ELPD) difference from the top-ranking model was within two times the standard error difference (negative values indicate a worse fit in comparison to the top-ranking model; see Appendix B). We selected the simplest model (e.g., lacking any interactions) to interpret when ELPD did not differ substantially, indicating that the added interaction did not improve model accuracy (see Appendix B).

To investigate trial outcome behavior (Q3a, Q3b), we fit a GLMM for each experiment with a logit-link function and the family specified as ‘categorical’ as there were three possible trial outcomes (completed, uncompleted, aborted). We set the reference category to ‘completed’

trials for both models. Choice, delay, and their interaction were our main test predictors for both experiments, including the same control predictors as the choice behavior models (monkey identity, choice position, reward accumulated, trial number; Table 3.2). As random effects, we included the same structures as described for the choice behavior models. Since the two monkeys may have responded differently to the task again, we considered whether including the interaction of monkey identity with delay and/or choice improved model fit using LOO cross-validation (Vehtari et al., 2017). Therefore, we compared nine models for each trial outcome analysis, beginning with a three-way interaction between delay, choice, and monkey identity as compared to the different combinations of reduced interactions and main effects (see Appendix B).

Model estimates are reported in Table 3.3 and Table 3.4 as the mean of the posterior distribution with 95 % credible intervals (CI). We calculated the proportion of posterior samples that fell on the same side of 0 as the mean (Pr) to aid in the interpretation of whether the predictor variables substantially affected choice or trial outcome behavior. As the Pr ranges from 0.5 to 1.0, a Pr of 1.0 indicates the direction (negative or positive) of a predictor's effect, whereas a Pr of 0.5 indicates an effect centered around 0 (i.e., no effect on response variable).

Table 3.2: Experimental variables, their definitions, type, levels, or range of the variable, and what model(s) they are present in. Choice models 1 and 2 refer to the analyses of choice behavior for the fixed and generalized stimuli experiments, respectively. Trial outcome (TO) models 1 and 2 refer to the analyses of trial outcome for the fixed and generalized stimuli experiments respectively.

Variable	Definition	Levels or Range [Unit]	Type	Models
Trial outcome	Whether the monkey completed, uncompleted, or aborted the trial (factor).	completed, uncompleted, aborted	Response	TO 1, TO 2
Choice ^a	Reward option chosen by the monkey (factor).	high, low	Response	Choice 1, Choice 2
			Test	TO 1, TO 2
Delay (fixed stim. exp.)	Duration of time the monkey waited to received reward (covariate).	1 - 600 [ms]	Test	Choice 1, TO 1
Delay (generalized stim. exp.)	Category of delay the monkey waited to receive reward (factor).	2, 6, 10 [min]	Test	Choice 2, TO 2
Monkey ID	Identity of the monkey (factor).	han, der	Control	All
Choice position	Position of the choice on the touchscreen (factor).	left, right	Control	All
Reward accumulated	Amount of reward received within the session (covariate).	0.15 - 667.85 [ml]	Control	All
Trial number	Trial number of within a session (covariate).	1 - 222	Control	All
Stimulus set	Identity of stimulus pairs in the generalized stimuli experiment.	2a, 2b, 2c, 2d, 6a, 6b, 6c, 6d, 10a, 10b, 10c, 10d	Random effect	Choice 2, TO 2
Session ID ^b	Unique dates that were tested.	1 - 31	Random effect	All
Information state ^c	Whether the monkey had experienced the consequences of both options (informed) or not (uninformed).	informed, uninformed	Other	–

^aChoice was a response or test variable depending on the model tested.; ^bSession ID was the random effect for models choice 1 and TO 1. Session ID nested within stimulus set was the random effect for models choice 2 and TO 2.; ^cAnalyses of the generalized stimuli experiment used the data of informed trials as the monkeys had experienced the consequences of both choices at least once.

3.3 Results

We determined if adult male rhesus macaques can learn to associate abstract stimuli with different quantities of positive reinforcement (high or low reward) despite reward delivery occurring up to 10 min later. In our task, the monkeys were presented with a choice between two stimuli, which were followed by the shape of the chosen stimulus that expanded linearly with delay (i.e., “expanding clock”). Once the delay expired, the stimulus illuminated and the monkeys needed to touch it to receive the corresponding fluid reward. We increased delay in a stepwise fashion up for the same stimulus set (fixed stimuli experiment) to determine if the monkeys would continue to wait to engage the final stimulus. In a second experiment (generalized stimuli experiment), we set delay to discrete durations (2, 6, or 10 min) and presented the monkeys with novel stimulus sets to determine if effects were independent of incremental training that may have occurred in the first experiment.

Overall, we conducted 8 test sessions per monkey in the fixed stimuli experiment and 14 or 15 sessions per monkey in the generalized stimuli experiment, resulting in 2648 trials overall (fixed stimuli experiment: 1044; generalized stimuli experiment: 1604 trials). In our analyses, we did not include trials where the monkeys made no decision (fixed stimuli experiment: 2.3 – 8.3 %; generalized stimuli experiment: 0.0 – 3.3 %) as these trials were few and were not meaningful to our research questions. As the monkeys were presented with 12 new sets of stimulus pairs in the generalized stimuli experiment, the monkeys did not have complete information about the stimuli until they had experienced the consequences of each (i.e., information state). In this respect, 21.0 % of trials occurred before the monkeys had experienced the consequences of both stimuli (i.e., uninformed; Table 3.2). Generally, we found that choices for the high reward stimulus of completed trials increased once the monkeys became informed of the consequences of both stimuli (see Appendix B). Therefore, we did not include uninformed trials in our analyses of the generalized stimuli experiment (Q3a, Q3b). Overall, 986 trials were entered into the analyses of the fixed stimuli experiment (Q2a, Q3a) and 1247 trials were entered into the analyses of the generalized stimuli experiment (Q2b, Q3b).

3.3.1 Q1: Do the monkeys learn to wait up to 10 min for the consequence of a choice between abstract stimuli (fixed stimuli experiment)?

To test whether monkeys could learn to wait up to 10 min for the consequence of a choice between abstract stimuli, over the course of the fixed stimuli experiment, we incrementally increased the time the monkeys had to wait to receive the consequence of their choice. Both monkeys were able to reach 10 min of delay within the task by test session 14 (monkey D after 524 trials; monkey H after 430 trials; Figure 3.2), hence, can associate abstract stimuli with their respective consequences occurring 10 min later.

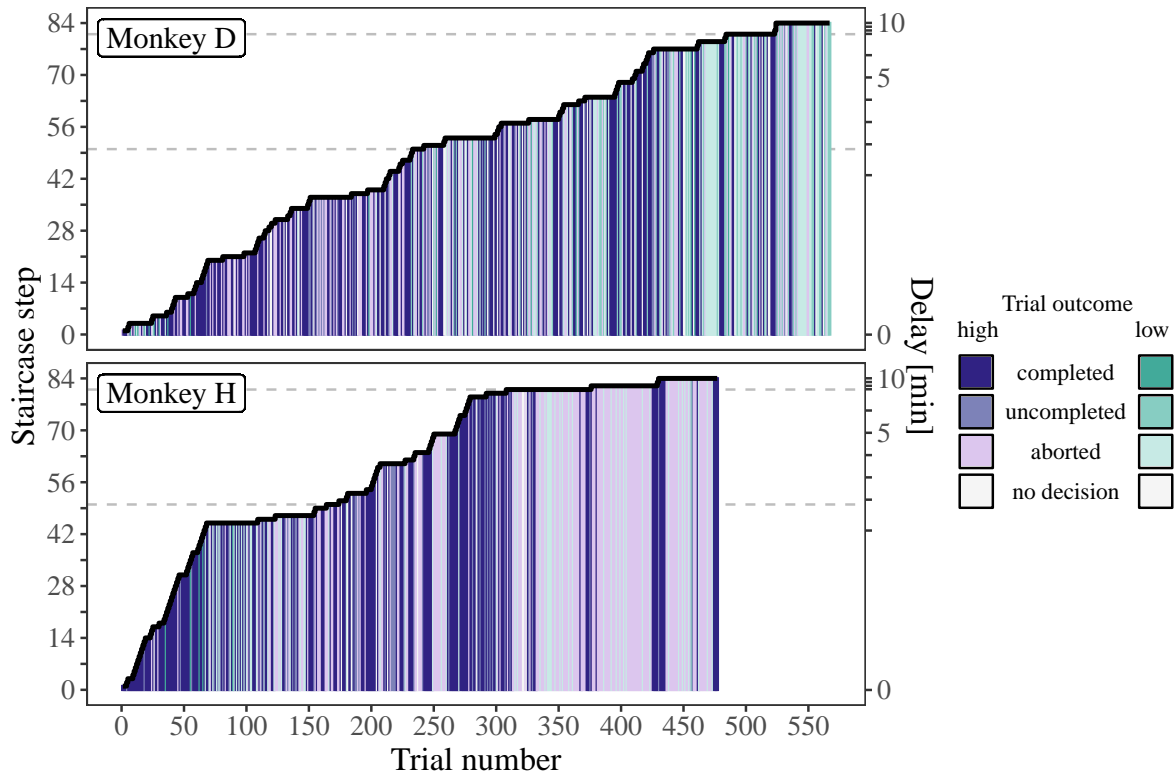


Figure 3.2: Staircase progress for the fixed stimuli experiment. Each trial is colored by the monkey's choice and its trial outcome. Dashed lines indicate a change in the step size (i.e., delay increase). Delay is scaled on the right Y-axis to indicate how it relates to staircase step.

3.3.2 Q2a: Is the monkeys' choice behavior modulated by increasing delay (fixed stimuli experiment; choice model 1)?

To determine if the monkeys developed a preference for the stimulus associated with high reward in the fixed stimuli experiment, we considered the effect of delay and monkey identity on choice behavior in a binomial GLMM. There was little evidence that the monkeys differed substantially in their choices as delay increased in the fixed stimuli experiment. LOO cross-validation indicated that the model containing an interaction between delay and monkey identity did not perform substantially better than the model lacking this interaction (ELPD difference: 0.33 ± 2.44 ; see Appendix B). Generally, the monkeys chose the high reward stimulus more often than the low reward stimulus (monkey D: 74.6 % of trials; monkey H: 85.4 % of trials; Figure 3.3). Additionally, we found strong evidence that the probability the monkeys made high reward choices declined as delay increased (Table 3.3; Figure 3.3).

3.3.3 Q2b: Is the monkeys' choice behavior modulated by increasing delay, despite stimulus pairs being novel (generalized stimuli experiment; choice model 2)?

We conducted the generalized stimuli experiment to check whether the results of the fixed stimuli experiment can be explained by a possible training effect due to delay increasing via

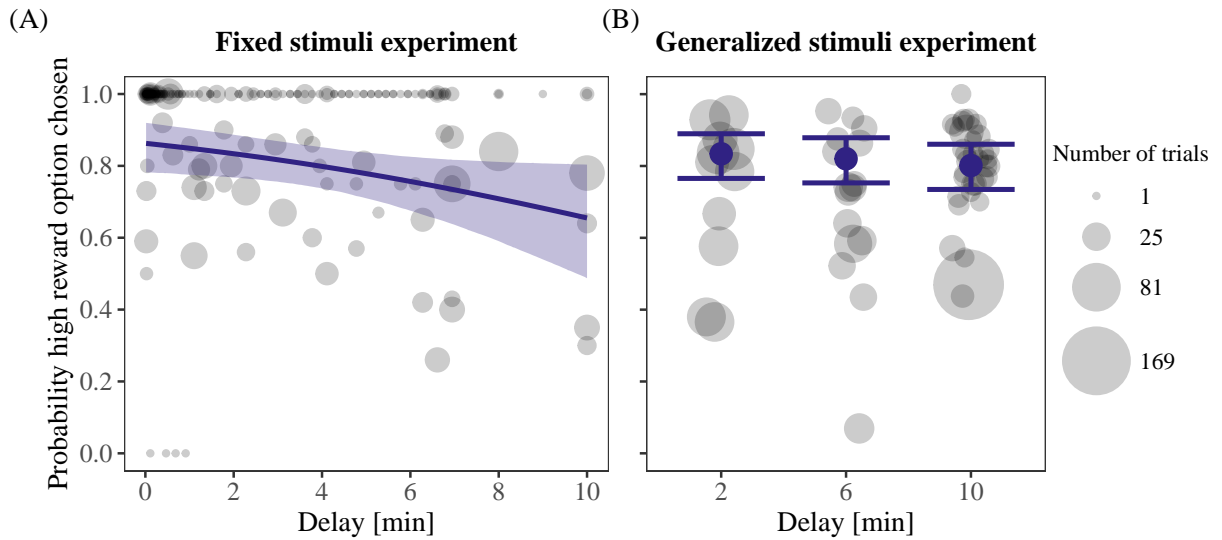


Figure 3.3: Exploring choice behavior during the fixed and generalized stimuli experiments. (a) Proportion of choices made for the high reward stimulus as delay increased during the fixed stimuli experiment. (b) Proportion of choices made for the high reward stimulus for the three set delays (2, 6, and 10 min) during the generalized stimuli experiment. Point size reflects the total number of trials conducted by each monkey, delay, and session in plot A, plus each stimulus set in plot B (range: 1 to 179 trials; see legend). The position of each point represents the proportion of trials that the high reward stimulus was chosen. Dark colored lines (plot A) and dark colored points (plot B) indicate the probability estimates of the models. Light bands (plot A) and whiskers (plot B) represent the 95 % credible intervals. Mean model probability estimates and credible intervals were calculated from models with all other variables are at their mean (factors dummy coded).

an experimental staircase. Therefore, we modified the task so that the delay was set to a predetermined duration (2, 6, or 10 min). For each delay, we tested four novel stimulus sets. To determine if the monkeys exhibited a preference for the stimulus associated with high reward in the generalized stimuli experiment across all trial outcomes, we considered the effect of delay and monkey identity on choice behavior in a binomial GLMM. Despite the novelty of the stimulus sets, the monkeys chose the high reward stimulus more often than the low reward stimulus (monkey D: 68.7 % of trials; monkey H: 72.0 % of trials; Table 3.3; Figure 3.3). LOO cross-validation indicated that the model containing an interaction between delay and monkey identity did not differ substantially in performance than the model lacking this interaction (ELPD difference: 0.01 ± 0.94 ; see Appendix B). The lack of interaction suggests that the two monkeys did not differ substantially in their choice behavior as delay increased. In contrast to the fixed stimuli experiment, we did not find evidence that delay influenced choice behavior in the generalized stimuli experiment across monkeys as the probability to choose the high reward stimulus remained high across delays (Table 3.3; Figure 3.3).

Table 3.3: Output of the choice analyses. Session ID was included as a random effect in choice model 1 and session nested within stimulus set was included as a random effect in choice model 2.

		Est.	SD	Lower CI	Upper CI	Pr
Choice model 1: Is choice modulated by increasing delay?						
Intercept		0.81	0.20	0.42	1.19	1.00
Test predictor	Delay	-0.67	0.20	-1.06	-0.27	1.00
Control predictors	Monkey ID (han) ^a	0.99	0.30	0.36	1.56	1.00
	Choice position (right) ^b	0.35	0.20	-0.04	0.77	0.96
	Trial number	-0.11	0.19	-0.50	0.25	0.71
	Reward accumulated	0.11	0.14	-0.15	0.39	0.78
Choice model 2: Is choice modulated by increasing delay, despite stimuli being <i>novel</i>?						
Intercept		1.07	0.23	0.63	1.52	1.00
Test predictor	Delay (6 min) ^c	-0.17	0.27	-0.71	0.37	0.73
	Delay (10 min) ^c	-0.25	0.28	-0.78	0.32	0.81
Control predictors	Monkey ID (han) ^a	0.54	0.21	0.15	0.95	1.00
	Choice position (right) ^b	0.97	0.22	0.56	1.43	1.00
	Trial number	-0.13	0.14	-0.40	0.14	0.82
	Reward accumulated	0.32	0.12	0.08	0.57	1.00

Est.: estimate, slope of the predictor. SD: standard deviation of the estimate. CI: 95 % credible interval. Pr: proportion of the posterior samples that fall on the same side of 0 as the mean.

^ader was the reference level for monkey ID in both models.; ^bLeft was the reference level for choice position in both models.; ^c2 min was the reference level for delay in the generalized stimuli experiment.

3.3.4 Q3a: Is the monkeys' trial outcome behavior influenced by their choice or/and delay (fixed stimuli experiment; trial outcome model 1)?

Generally, the monkeys completed (monkey D: 44.3 %; monkey H: 41.5 %) and aborted trials (monkey D: 33.2 %; monkey H: 47.0 %) more often than incompleting trials (monkey D: 14.3 %; monkey H: 9.2 %) in the fixed stimuli experiment (monkey D and H made no decisions on 8.3 and 2.3 % of trials respectively). To investigate if and how the monkeys may have treated trials of each stimulus differently in this experiment, we considered the effect of delay, choice, and monkey identity on trial outcome behavior (whether the monkeys completed, uncompleted, or aborted a trial) in a categorical GLMM. The top-ranking model for the trial outcome analysis of the fixed stimuli experiment included two 2-way interactions: choice interacting with delay, and choice interacting with monkey identity (see Appendix B). Although this model did not differ in performance from the second and third ranked models as determined by LOO cross-validation (ELPD difference with second ranked: 0.62 ± 0.69 ; ELPD difference with third ranked: 0.97 ± 0.76 ; see Appendix B), it was the simplest in its interaction structure. According to this model, monkeys were generally more likely to complete trials when the high reward stimulus was chosen than when the low reward was chosen (Figure 3.4). Moreover, the probability to complete trials declined with increasing delay, independent of the stimulus chosen (Figure 3.4). In comparison to completed trials, the monkeys' likelihood to uncomplete trials was lower for both stimuli (Figure 3.4), but increased for low, but not high, reward choices with increased delay (Table 3.4; Figure 3.4). With respects to aborted trials, the monkeys exhibited the opposite pattern to completed trials, where the probability to abort was generally higher for low rewarded trials than high reward trials (Figure 3.4). We found strong evidence that the monkeys' probability to abort increased with increasing delay for low reward trials and weak evidence of such a relationship for high reward trials (Table 3.4; Figure 3.4).

While both monkeys had a higher probability to complete trials when the high reward stimulus was chosen than when the low reward stimulus was chosen, the two differed with respects to how they uncompleted and aborted low reward trials (Table 3.4). Monkey D uncompleted low reward choice trials more frequently than high reward, albeit he exhibited more variation for low reward choice trials (Table 3.4). In contrast, monkey H uncompleted trials of both stimuli at a similar low frequency (Table 3.4). Generally, both monkeys had a higher probability to abort trials than to uncomplete them and exhibited a similar pattern where he aborted low reward choice trials more frequently than high reward choice trials (Table 3.4). However, the data suggest that monkey H exhibited a stronger difference in the probability to abort trials for each stimulus than monkey D (Table 3.4). See Appendix B for the plotted data.

3.3.5 Q3b: Is the monkeys' trial outcome differ behavior influenced by their choice or/and delay, despite stimulus pairs being novel (generalized stimuli experiment; trial outcome model 2)?

Across trials in the generalized stimuli experiment, the monkeys completed (monkey D: 56.8 %; monkey H: 50.3 %) trials more often than uncompleting (monkey D: 21.8 %; monkey H: 14.4 %) or aborting (monkey D: 18.2 %; monkey H: 34.9 %) them (monkey D and H made no decisions on 3.3 and 0.0 % of trials, respectively). To investigate trial outcome behavior in the generalized stimuli experiment further, we considered the effect of delay, choice, and monkey identity (whether the monkeys completed, uncompleted, or aborted a trial) in a categorical GLMM. All possible combinations of choice, delay, and monkey identity as interactions and/or as main effects did not differ substantially in performance (largest ELPD difference: -2.08 ± 1.67 ; see Appendix B). Therefore, we interpreted the simplest model comprised of these variables as main effects alone. Generally, there was not sufficient evidence that the monkeys differed in their trial outcome behavior with respects to their choices and delay in the generalized stimuli experiment.

In contrast to the fixed stimuli experiment, delay did not influence trial outcome behavior in the generalized stimuli experiment (Table 3.4). Which stimulus the monkeys chose had an effect whether they completed or uncompleted trials, but not whether they aborted trials (Table 3.4; Figure 3.4). Specifically, the monkeys had a higher probability to complete high reward than low reward trials, whereas they were less likely to uncomplete high reward trials than low reward trials (Table 3.4; Figure 3.4).

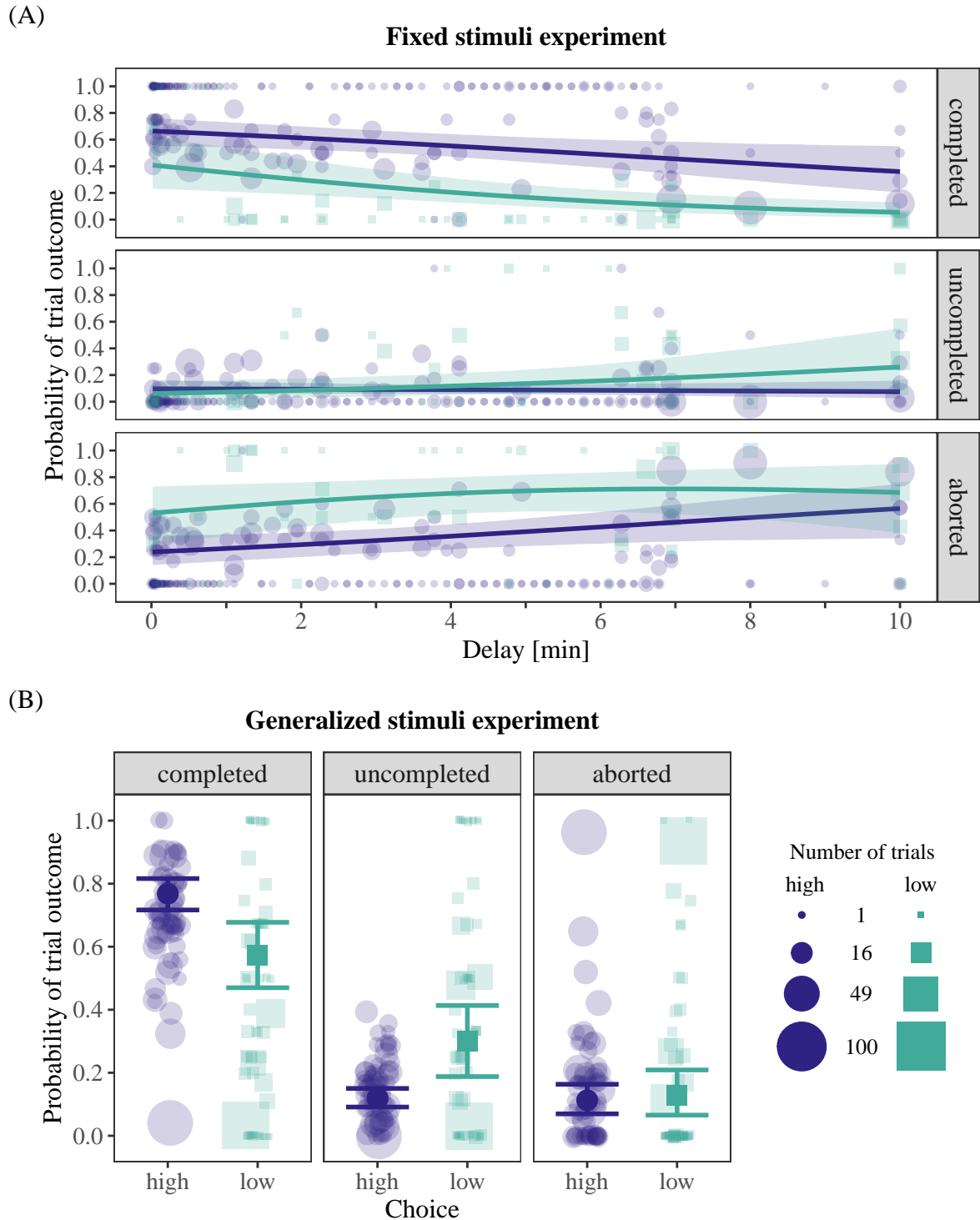


Figure 3.4: Exploring trial outcome behavior during the constant and generalized stimuli experiments. (a) Probability of each trial outcome (completed, uncompleted, aborted) by choice (low reward, high reward) as delay increased during the fixed stimuli experiment. (b) Probability of each trial outcome by choice during the generalized stimuli experiment (no effect of delay found). Point size reflects the total number of trials conducted by each monkey, delay, session, and choice in plot A, plus stimulus set in plot B (range: 1 to 94 trials; see legend). The position of each point represents the proportion of trials that were completed, uncompleted, or aborted. Dark colored lines (plot A) and dark colored points (plot B) indicate the model probability estimates. Light bands (plot A) and whiskers (plot B) represent the 95 % credible intervals. Model probability estimates and credible intervals calculated from a model run with all other variables are at their mean (factors dummy coded).

Table 3.4: Output of the trial outcome analyses. Session ID was included as a random effect in choice model 1 and session nested within stimulus set was included as a random effect in choice model 2.

			Est.	SD	Lower CI	Upper CI	Pr
Trial outcome model 1: Is trial outcome behavior influenced by choice or/and increasing delay?							
Uncompleted							
Intercept			-1.95	0.24	-2.43	-1.49	1.00
Test predictors	Delay	*High reward	0.12	0.22	-0.30	0.54	0.71
		*Low reward	1.16	0.33	0.52	1.82	1.00
	Monkey der	*High reward	-1.95	0.24	-2.43	-1.49	1.00
		*Low reward	-0.43	0.39	-1.30	0.25	0.88
	Monkey han	*High reward	-1.98	0.29	-2.57	-1.41	1.00
		*Low reward	-2.49	0.52	-3.50	-1.45	1.00
Control predictors	Choice position (right) ^a		0.25	0.24	-0.24	0.71	0.85
	Trial number		-0.12	0.21	-0.54	0.30	0.72
	Reward accumulated		0.57	0.17	0.24	0.90	1.00
Aborted							
Intercept			-0.34	0.20	-0.75	0.04	0.96
Test predictors	Delay	*High reward	0.35	0.20	-0.04	0.73	0.96
		*Low reward	0.80	0.27	0.27	1.34	1.00
	Monkey der	*High reward	-0.34	0.20	-0.75	0.04	0.96
		*Low reward	0.81	0.28	0.25	1.35	0.99
	Monkey han	*High reward	-0.32	0.34	-1.03	0.34	0.84
		*Low reward	0.77	0.52	-0.32	1.72	0.93
Control predictors	Choice position (right) ^a		-0.39	0.18	-0.74	-0.03	0.98
	Trial number		0.26	0.20	-0.13	0.66	0.92
	Reward accumulated		-0.30	0.20	-0.71	0.09	0.93

		Est.	SD	Lower CI	Upper CI	Pr
Trial outcome model 2: Is trial outcome behavior influenced by choice or/and increasing delay, despite stimuli being novel?						
Uncompleted						
Intercept		-1.51	0.23	-1.97	-1.06	1.00
Test predictors	Choice (low reward) ^b	1.23	0.28	0.63	1.74	1.00
	Delay (6 min) ^c	0.02	0.27	-0.52	0.54	0.52
	Delay (10 min) ^c	-0.22	0.29	-0.80	0.34	0.78
Control predictors	Monkey ID (han) ^d	-0.61	0.21	-1.03	-0.21	1.00
	Choice position (right) ^a	0.05	0.18	-0.30	0.40	0.63
	Trial number	-0.02	0.18	-0.39	0.34	0.54
	Reward accumulated	0.72	0.12	0.49	0.97	1.00
Aborted						
Intercept		-1.61	0.26	-2.14	-1.09	1.00
Test predictors	Choice (low reward) ^b	0.35	0.26	-0.20	0.84	0.90
	Delay (6 min) ^c	-0.31	0.32	-0.94	0.34	0.83
	Delay (10 min) ^c	-0.28	0.35	-0.97	0.39	0.79
Control predictors	Monkey ID (han) ^d	-0.24	0.35	-0.93	0.45	0.76
	Choice position (right) ^a	-0.39	0.20	-0.81	0.01	0.97
	Trial number	0.17	0.24	-0.29	0.65	0.77
	Reward accumulated	-0.50	0.17	-0.84	-0.16	1.00

Est.: estimate, slope of the predictor. SD: standard deviation of the estimate. CI: 95 % credible interval. Pr: proportion of the posterior samples that fall on the same side of 0 as the mean.

^aLeft was the reference level for choice position.; ^bHigh reward was the reference level for choice.; ^c2 min was the reference level for delay.; ^dder was the reference level for monkey ID.

3.4 Discussion

Our study aimed to determine if rhesus macaques can learn to discriminate between abstract stimuli, even if the associated positive reinforcement is delayed by as much as 10 minutes. Both monkeys successfully reached 10 minutes of delay between their choice and its respective consequences in our task. Their strong preference for the high value stimulus, documents that they have learned to discriminate between the two abstract stimuli despite the long-delayed feedback. The monkeys also continued to engage in the task. These two elements demonstrate long-delay learning capabilities in macaques. Our findings were consistent across both the fixed and generalized stimuli experiments, where the delay was increased in a stepwise fashion for the same stimulus set and set to discrete durations for novel stimulus sets, respectively.

Altogether, our study suggests that monkeys are capable of long-delay learning with abstract stimuli and positive reinforcement, a condition that few studies have explicitly tested. Although Ferster & Hammer (1965) demonstrated that macaques and baboons could be conditioned to respond to keys for positive reinforcement in relation to a delay period, no choice was offered in their study. Tasks involving choices between options, like ours, can indicate preferences, which arguably demonstrates an active decision and understanding of the association between abstract stimuli and its delayed reinforcement. For example, a study (D'Amato et al., 1981) in capuchin monkeys failed to show that this species is able to use abstract visual stimuli for delayed reinforcement in a T-maze task, but navigation signals may have interfered. When navigation signals were the only modality necessary to distinguish, capuchins seem capable of long-delay learning (D'Amato & Puopolo, 1981; Safarjan & D'Amato, 1981). Similarly, we find that long-delay learning from abstract stimuli is feasible when the influence of other modalities is minimized.

Our study shows long-delay learning capabilities for delays on the order of minutes using positive outcomes. This is noteworthy because similar studies typically use aversive outcomes. Much of aversion research involves experiments where the animals are presented with a particular stimulus (e.g., food item) followed by the administration of a substance that induces delayed gastrointestinal discomfort (reviewed in Bernstein, 1999). For example, Japanese macaques (*M. fuscata*) that received an injection of an illness-inducing agent avoided an associated food after one trial, whereas those receiving an injection of saline continued to consume the food (Matsuzawa et al., 1983). It is important to note that aversion learning evolved as means to avoid poisoning (reviewed in Bernstein, 1999; LeDoux, 2003). Thus, there is likely more selective pressure to learn associated stimuli than those associated with different positive reinforcement. Further exploration is needed to understand if delay lengths in a choice-based delayed positive reinforcement tasks can be extended to hours as is possible in aversion studies (reviewed in Bernstein, 1999).

Our data suggest that long-delay learning may be facilitated when delays are more predictable. We show that delay influenced choice and trial outcome behavior when it is continuously

increasing (fixed stimuli experiment), but does not when it is set to discrete durations (generalized stimuli experiment). Specifically, the monkeys chose the high reward stimulus and completed trials less frequently as delay increased in the fixed stimuli experiment. However, such behavior may have been due to the nature of the delay schedule and changes in step size (increased from 10-second to 1-minute increments once 7 minutes of delay was reached). This regularly changing delay may have caused the monkeys to try out the other option more often and be generally less engaged with the task. A similar reduction in engagement in a delayed positive reinforcement task was reported by Ferster & Hammer (1965), who observed that the overall rate of responding by rhesus macaques decreased as delay was increased. In contrast, the set delay durations in our generalized stimuli experiment possibly allowed the monkeys to become familiar with the timing of the delay despite the stimulus sets changing more frequently.

We show that shaping (e.g., incremental increases in delay) is not a prerequisite for long-delay learning. The monkeys were able to discover and reliably select the higher value stimulus when stimulus sets were novel in the generalized stimuli experiment. Generally, the principles of animal training recommend that more complex associations between a stimulus and its consequence, such as long-delay, necessitate shaping in steps or cycles (Kurland & St. Peter, 2022). In contrast, our findings indicate that long-delay learning was not predicated by incremental increases in delay beginning from short durations. Notably, the shortest delay duration in our generalized stimulus experiment was 2 minutes. Similar evidence was found by Ferster & Hammer (1965), who showed macaques engaged in an operant task despite a 24-hour delay when substantial food reward was delivered. Our study extends these previous findings by incorporating multiple options that the monkeys could choose between.

Stimulus quality influences choice and trial outcome behavior. Trials following the selection of the high or low reward stimulus were treated differently by the monkeys in our study. High reward trials were completed more frequently than low reward trials, whereas low reward trials were aborted and/or uncompleted (depending on the experiment) more often than high reward trials. Such differential treatment of the stimuli suggests that the monkeys maintained their commitment to a high reward decision and retained information about the quality of the stimuli (i.e., understood the stimuli associations with reward, Hayden, 2016). In contrast, the monkeys' higher likelihood to abort and/or uncomplete low reward choice trials suggests that they were more selective about continuing to engage with these trials as the reward benefits were substantially lower.

We observed that the monkeys infrequently aborted trials early. This observation documents that impulsive responses were rare. Such behavior contrasts to other studies testing delay tolerance capabilities in non-human primates using the intertemporal task (i.e., delay choice), where subjects are offered a choice between receiving a small reward sooner and a larger reward later. These studies suggest delay tolerance capabilities of several seconds in new world species (e.g., Stevens et al., 2005) to up to several minutes in great apes (e.g., Rosati et al., 2007).

Accumulating evidence suggests that non-human primates may actually lack an understanding of the intertemporal task's temporal structure due to the time buffer generally added to equalize the trial lengths to make the larger reward later the more optimal option (Blanchard et al., 2013; e.g., Blanchard & Hayden, 2015). This misunderstanding may be due to the lack of explicit time information (Blanchard et al., 2013).

Impulsive behavior was likely reduced by the use of abstract stimuli instead of edible stimuli. Because the use of edible stimuli requires impulse control, subjects must suppress natural instincts to reach towards which ever choice they see first (e.g., Schmitt & Fischer, 2011) or for the larger array of rewards (e.g., Genty et al., 2012). Comparative cognition studies testing the intertemporal task have typically presented edible rewards as the options by experimenters (e.g., Rosati et al., 2007; Stevens et al., 2005). These instincts may increase spurious conclusions about delay tolerance capabilities (Paglieri et al., 2015). Additionally, the presence of the experimenter may confound natural behavior due to their previous interactions (Sato et al., 2021; Schmitt et al., 2014). For example, experimenters are generally associated with training, where a behavior is performed for a reward. Thus, their mere presence may unintentionally suggest that an operant response is necessary sooner rather than later.

The monkeys' choice behavior does raise the question, why our monkeys choose the low reward stimulus at all once informed of the consequences of each stimulus. The optimal strategy would be to exclusively choose the high reward stimulus once its value becomes apparent. Choosing the low reward stimulus at all during the fixed stimuli experiment is puzzling as the same stimulus set was tested throughout. Thus, the monkeys were very familiar with the associated consequences of each stimulus. One possible explanation is that increase in delay may have triggered the monkeys to test out if the reward association with the low reward stimulus had changed. In the generalized stimuli experiment, the monkeys were less familiar with these stimulus sets (i.e., each set tested over fewer days than in the fixed stimuli experiment). Therefore, the monkeys' lack of experience with novel stimulus sets likely accounts for more low reward stimulus choices than expected.

Often animals are not given information about how long they need to wait in tasks testing delay tolerance. Knowing about how long one needs to wait might, however, be helpful for determining whether it is worth it to stay engaged. Our "expanding clock" was an informative secondary reinforcer as it marked and tracked the delay period and served as a memory of the stimulus choice the monkey made. Studies have found that learning is enhanced when there is link or chain of events between a stimulus and its outcome (reviewed in Lattal, 2010). We aided the perception of this chain of events by testing the monkeys in their home-cage environment, avoiding the distractions that may occur due to transport and training in a primate chair (e.g., Pfefferle et al., 2018), which could impede with learning (e.g., Costa et al., 2022).

For future applications, our findings provide a proof-of-concept that rhesus macaques are capable of long-delay learning. Testing for the presence and limits of such capabilities offers further insights into the learning processes of animals. Such capabilities may also allow

researchers and caretakers to train animals to associate simple stimuli with complex options or events that may involve delay. Signaling husbandry feeding times by a reliable auditory stimulus, for example, can disassociate feeding with out of sight caretaker activity and thereby improve behavioral measures of welfare in rhesus macaques (Gottlieb et al., 2013b). We are actively using the results of this study in a project offering animals choices between complex procedures, such as performing a cognitive task in their home environment versus a laboratory environment, to provide a window into their subjective experience. Such knowledge of animal preferences can help guide and optimize captive care and research practices and thereby improve animal welfare (Schapiro & Lambeth, 2007).

In summary, we found that adult male rhesus macaques can tolerate lengthy delays (\geq 10 minutes) when learning stimulus-reward associations, even when the initial choice is an abstract stimulus. Beyond this insight into the cognitive flexibility of rhesus macaques, our findings facilitate the development and application of complex cognitive paradigms and welfare assessments, necessitating such long-delays before animals are rewarded.

Chapter 4

The dot-probe attention bias task as a method to assess psychological wellbeing after anesthesia: A study with adult female long-tailed macaques (*Macaca fascicularis*)

Lauren C. Cassidy^{a,b}, Emily J. Bethell^{c,d}, Ralf R. Brockhausen^a, Susann Boretius^{b,e}, Stefan Treue^{a,b}, Dana Pfefferle^{a,b}

^aWelfare and Cognition Group, Cognitive Neuroscience Laboratory, German Primate Center - Leibniz Institute for Primate Research, Göttingen, Germany; ^bLeibniz-ScienceCampus Primate Cognition, German Primate Center & University of Göttingen, Göttingen, Germany; ^cLiverpool John Moores University, Research Centre in Evolutionary Anthropology and Palaeoecology, Liverpool, England; ^dLiverpool John Moores University, Research Centre in Brain and Behaviour, Liverpool, England; ^eFunctional Imaging Laboratory, German Primate Center - Leibniz Institute for Primate Research, Göttingen, Germany

This article was published (online: 16 December 2021) in *European Surgical Research* (2021) as a part of a special issue on severity assessment in laboratory animals. DOI: 10.1159/000521440. Permission to reprint the article within this thesis was received from S. Karger AG, Basel on behalf of *European Surgical Research* on 30 June 2022.

Contribution to the field

Affect-mediated changes in cognition can offer insight into the psychological states of animals, but methods detecting such changes require validation. In the following study, we determine if a dot-probe task can detect changes in the attention bias — the tendency to attend to one type of information over another — of adult female long-tailed macaques in relation to experiencing prolonged anesthesia. We found evidence that the pattern of attention bias changed from a vigilance to threatening stimuli to avoidance on the day immediately following prolonged anesthesia. With refinements and further validation, the dot-probe attention bias task has the potential to be a powerful animal welfare assessment tool and offer deeper insight into the psychological experiences of animals.

Author contributions

Lauren Cassidy, Emily Bethell, Dana Pfefferle, and Ralf Brockhausen contributed to the conception and design of the study. Ralf Brockhausen programmed the task. Susann Boretius provided access to the monkeys and conducted the anesthesia study. Stefan Treue provided the test systems and infrastructure support. Lauren Cassidy trained the monkeys and collected the data. Lauren Cassidy performed the statistical analyses, with advice from Emily Bethell and Dana Pfefferle. Lauren Cassidy, Emily Bethell, and Dana Pfefferle interpreted the data. Lauren Cassidy wrote the manuscript, with revision by Dana Pfefferle, Emily Bethell, and Stefan Treue. All authors read, revised, and approved the submitted version of the manuscript.

Acknowledgments

We thank the two anonymous reviewers for their constructive feedback on the manuscript, which helped improve the quality of the article substantially. We would also like to thank the German Primate Center animal care and veterinary staff for taking care of the monkeys.

Statement of ethics

See in publication.

Conflict of interest statement

This research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding sources

This study was supported by a Leibniz-ScienceCampus Primate Cognition Seed Fund for DP and SB (<https://www.primate-cognition.eu/en/funding-measures/seed-funds.html>, Grant No. LSC-2017-01SF), a Leibniz-ScienceCampus Primate Cognition Incoming Grant to DP (<https://www.primate-cognition.eu/en/funding-measures/incoming-grants.html>, Grant No. LSC- 2018-01-IG), and the German Research Foundation Research unit 2591 “Severity assessment in animal-based research” assigned to ST, Alexander Gail, and DP (Grant No.: TR 447/5-1/2, GA 1475/6-1/2, PF 659/5-2). Additionally, EB and DP received funding from the European Cooperation in Science and Technology from the Short-Term Scientific Mission program (2017: ECOST-STSM-CA15131-010317-084470; 2019: ECOST-STSM-CA15131-45030).

Supplementary material and data

See Appendix C for the associated supplementary material. The data supporting this article can be found in the online repository, *Göttingen Research Online*. DOI: 10.25625/ATAAGM.

The Dot-Probe Attention Bias Task as a Method to Assess Psychological Well-Being after Anesthesia: A Study with Adult Female Long-Tailed Macaques (*Macaca fascicularis*)

Lauren C. Cassidy^{a, b} Emily J. Bethell^{c, d} Ralf R. Brockhausen^a
Susann Boretius^{b, e} Stefan Treue^{a, b} Dana Pfefferle^{a, b}

^aWelfare and Cognition Group, Cognitive Neuroscience Laboratory, German Primate Center–Leibniz Institute for Primate Research, Goettingen, Germany; ^bLeibniz-Science Campus Primate Cognition, German Primate Center, University of Goettingen, Goettingen, Germany; ^cLiverpool John Moores University, Research Centre in Evolutionary Anthropology and Palaeoecology, Liverpool, UK; ^dLiverpool John Moores University, Research Centre in Brain and Behaviour, Liverpool, UK; ^eFunctional Imaging Laboratory, German Primate Center–Leibniz Institute for Primate Research, Goettingen, Germany

Keywords

Dot-probe task · Attention bias · Affect · Emotion · Anesthesia

Abstract

Understanding the impact routine research and laboratory procedures have on animals is crucial to improving their well-being and to the success and reproducibility of the research they are involved in. Cognitive measures of welfare offer insight into animals' internal psychological state, but require validation. Attention bias – the tendency to attend to one type of information over another – is a cognitive phenomenon documented in humans and animals that is known to be modulated by affective state (i.e., emotions). Hence, changes in attention bias may offer researchers a deeper perspective of their animals' psychological well-being. The dot-probe task is an established method for quantifying attention bias in humans (by measuring reaction time to a dot-probe replacing pairs of stimuli), but has yet to be validated in animals. We developed a dot-probe task for long-tailed macaques (*Macaca fascicularis*) to determine if the task can

detect changes in attention bias following anesthesia, a context known to modulate attention and trigger physiological arousal in macaques. Our task included the following features: stimulus pairs of threatening and neutral facial expressions of conspecifics and their scrambled counterparts, two stimuli durations (100 and 1,000 ms), and counterbalancing of the dot-probe's position on the touchscreen (left and right) and location relative to the threatening stimulus. We tested 8 group-housed adult females on different days relative to being anesthetized (baseline and 1-, 3-, 7-, and 14-days after). At baseline, monkeys were vigilant to threatening content when stimulus pairs were presented for 100 ms, but not 1,000 ms. On the day immediately following anesthesia, we found evidence that attention bias changed to an avoidance of threatening content. Attention bias returned to threat vigilance by the third day postanesthesia and remained so up to the last day of testing (14-days after anesthesia). We also found that attention bias was independent of the type of stimuli pair (i.e., whole face vs. scrambled counterparts), suggesting that the scrambled stimuli retained aspects of the original stimuli. Nevertheless, whole faces were more salient to the monkeys as responses to these trials were

generally slower than to scrambled stimulus pairs. Overall, our study suggests it is feasible to detect changes in attention bias following anesthesia using the dot-probe task in nonhuman primates. Our results also reveal important aspects of stimulus preparation and experimental design.

© 2021 S. Karger AG, Basel

Introduction

Ensuring high standards of animal welfare is crucial for conducting ethical and reproducible biomedical and basic research. Ideally, methods for assessing welfare should be the objective and reflect changes to an animal's physiological or/and psychological well-being. Attention bias, the process of selectively attending to one type of information over another [1], is one cognitive process that may offer insight into animals' psychological well-being and affective state. Human attention biases are influenced by context, changes in physiology, mood, and intrinsic traits such as personality [2–4]. Attention bias tasks have found that humans, particularly for those with affective disorders like anxiety, preferentially attend to threatening information [5–9]. Given this evidence, tasks for detecting affect-mediated attention biases are being modified for animals to assess affect noninvasively [10].

Affect-mediated attention bias tasks are adapted for animals using biologically relevant stimuli that trigger innate responses such as gaze (e.g., [11]) or movement (e.g., [12]; reviewed in [10, 13]). Differences or changes in attention biases have been examined via trait affect (parrots, *Amazona amazonica*: [14]; rhesus macaques, *Macaca mulatta*: [15]), by manipulating state affect in individual animals (rhesus macaques: [16]) or groups of animals (mice, *Mus musculus*: [12]; starlings, *Sturnus vulgaris*: [17, 18]), and by comparing groups of animals administered with or without pharmacological anxiety drugs (cattle, *Bos taurus*: [19]; sheep, *Ovis aries*: [20]). For example, Bethell et al. [16] found that how male rhesus macaques attended to threatening and neutral facial expressions was modulated by the type of affect manipulation the males recently experienced. Specifically, males were more avoidant of threatening stimuli following a stressful veterinary procedure (i.e., health check) than after period of enrichment.

Looking-time experiments can help provide a complete picture of the different processing stages of attention: initial engagement, maintenance, and disengagement [21–23]. However, these experiments can be time-consuming (e.g., if video must be coded) or costly (if

eye-tracking equipment is required) [24, 25]. One alternative to looking-time experiments is the dot-probe task, which is sensitive to affect-mediated attention bias in humans (reviewed in [26–28]). In this task, participants are presented with a stimulus pair (e.g., facial expressions) simultaneously for a fixed duration. After the stimuli disappear, a “dot-probe” (i.e., neutral target) appears in the location of 1 stimulus, and the latency to touch this target is measured. Faster reaction times to the dot-probe indicate that attention was likely allocated toward the stimulus it replaced, whereas slower reactions suggest that attention shifted from another location, presumably the other stimulus. Manipulating the stimuli presentation duration allows researchers to capture the different stages of attention [29], which may reveal if participants show vigilance, avoidance, or a pattern of both to a particular stimulus [30, 31]. Stimulus pairs often consist of a stimulus with neutral content paired with another of high threatening content, with the latter capturing gaze automatically. Importantly – and different to other tasks measuring attention bias – these stimuli are task irrelevant (i.e., no trained reward contingencies) and may limit habituation due to their biological salience for the species being studied. Furthermore, animals can learn the dot-probe task easily as touching the dot-probe is the only rule they have to understand (e.g., [32]).

So far the detection of affect-mediated attention biases by the dot-probe task has been tested only in humans. A meta-analysis of studies investigating anxiety found that anxious participants were faster to react to dot-probes replacing the negative or threatening stimulus [26]. Similar findings have been reported for humans suffering from depression [27]. These findings attest that the dot-probe task is sensitive to trait affect and have provided the foundation for dot-probe studies testing context-driven attention changes. In this respect, dot-probe studies in humans involving affective manipulations have tested negatively valenced contexts ranging from acute stress induction (cold press test: [33]; mild contextual shock: [34]) to putatively severe, chronic stressors (rocket attack: [35]; combat deployment: [36]). How and if attention bias is modulated as detected by the dot-probe task may depend on gender (e.g., [33]) and level of stress exposure (e.g., [34–36]).

Given the supporting evidence from human studies, the dot-probe task shows potential for detecting affect-driven attention bias changes in other animals. Despite this potential, the dot-probe task has been implemented relatively rarely within the realm of animal cognition (reviewed in [10, 28]). Currently, dot-probe studies have

been conducted only in nonhuman primates (NHPs), focusing on comparing reactions to dot-probes replacing affective content to those replacing neutral content (bonobos, *Pan paniscus*: [32]; chimpanzees, *Pan troglodytes*: [37]; rhesus macaques: [38]; capuchins, *Sapujus apella*: [39]; summarized in online suppl. Table 5; see www.karger.com/doi/10.1159/000521440 for all online suppl. material). Yet no study to date has tested whether the task is also sensitive to changes in the affective state in these species, which bears potential as a welfare assessment method.

General anesthesia is a common and necessary procedure in veterinary medicine. Experiencing anesthesia is likely one of the strongest contexts that could influence affect in captivity, as it is a known physiological stressor (e.g., [40–42]). In addition to the anesthesia itself, associated processes, such as a social group separation for fasting, having the anesthetic applied, and waking up from surgery in isolation, likely cause additional physiological or/and psychological effects. We opportunistically tested the reliability and sensitivity of a dot-probe task for detecting changes in affect due to experiencing prolonged anesthesia in 8 adult female long-tailed macaques (*Macaca fascicularis*). Improving methodologies for assessing NHP psychological well-being is necessary as these species are crucial for the advancement of scientific and medical knowledge, treatments, and applications (reviewed in [43–45]). We tested the monkeys on the dot-probe task during a baseline test session, when no anesthesia had been administered at least 30 days prior, and at 4 test sessions following prolonged anesthesia (1-, 3-, 7-, and 14-days). Our dot-probe task for NHPs incorporated design features common among dot-probe studies (summarized in [28]). For this experiment, we assessed whether the dot-probe task detected (Q1) attention bias, (Q2) an affect-mediated change in attention bias following the anesthesia, and (Q3) when attention bias returned to baseline levels postanesthesia assuming a change occurred. We expected the dot-probe task to detect an attention bias to threat (specifically reacting more quickly to dot-probes replacing the aggressive face) for whole-face stimuli during the baseline test session. Additionally, we predicted that the dot-probe task would be able to detect a change in attention bias following anesthesia that would return to the monkeys' baseline levels of attention bias in the following days. We present our study as a guide for optimizing future studies as it is the first to implement the dot-probe in relation to an affect manipulation in an animal.

Materials and Methods

Study Subjects and Housing Facility

We conducted the study on 8 adult female long-tailed macaques living at the German Primate Center, Goettingen, Germany. The monkeys were housed in isosexual groups of 4 to 5 individuals with visual and auditory contact to other macaque groups. Age of the monkeys ranged from 6 to 19 years (mean \pm standard deviation: 11.3 ± 5.8 years) at the time point of the study. Housing consisted of a large indoor compartment with a 12-h light/dark cycle (from 07:00 to 19:00) connected by an elevated tunnel to an outdoor compartment where animals could experience natural light, temperature fluctuations, and wind, with visual access to the outdoors (living space exceeded the size requirements for macaques set by EU directive 2010/63/EU). Both areas were furnished with fixed and dynamic perching (e.g., raised platform and ropes), environmental enrichment (e.g., balls and cardboard), and carpeted with wood shavings. A flexible compartment (i.e., testing compartment) adjacent to the indoor living quarters was used for animal training, testing, temporary separation, and veterinary procedures (approximately 80 cm by 75 cm by 90 cm). Monkeys had access to water and monkey chow ad libitum and received fresh fruit and vegetables daily.

Experimental Testing Protocol

Our study took place between August 2017 and January 2018 and ran concurrently with another project investigating the effect of prolonged anesthesia on the brain using magnetic resonance imaging (see Statement of Ethics for permit information). Veterinarians regularly monitored the monkeys during the entire study. In preparation for anesthesia, monkeys were separated (but with visual, acoustic, and olfactory contact to their group members) and food removed the night before. Anesthesia was induced by a mixture of ketamine (mean \pm standard deviation: 8.0 ± 2.7 mg per kg) and medetomidine (mean \pm standard deviation: 0.02 ± 0.01 mg per kg), and maintained by isoflurane (0.8–1.7% in oxygen and ambient air) via an endotracheal tube and pressure-controlled active ventilation. The duration of isoflurane anesthesia ranged from 213 to 350 min (310 ± 42 min).

Following anesthesia, monkeys were kept separate overnight (but with visual, acoustic, and olfactory contact to their group members) for the purposes of recovery and observation. Monkeys were returned to their living quarters the following morning. Each monkey performed the dot-probe task once as a baseline, when no anesthesia had been administered at least 27 days prior, and at 4 time points following the anesthesia session: on average 1- (A + 1d), 3- (A + 3d), 7- (A + 7d), and 14-days (A + 14d) after. Baseline measurements occurred in a counterbalanced design. Five monkeys were tested at least 27 days before any monkeys in the group experienced prolonged anesthesia (range: 28–32 days). Three monkeys were tested at least 34 days after all anesthesia procedures occurred (range: 35–36 days). Due to the timing of baseline test sessions, we presume that these test sessions coincided with a period of comparatively low stress to the sessions immediately following prolonged anesthesia. It is possible daily fluctuations in stress may have occurred on the day of each test session due to social, environmental, or/and husbandry factors, for example. However, these influences are likely to be limited as we observed no increases in aggression (rare occurrence overall) or changes in hierarchy that may have indicated group instability (systematic be-

eral or contralateral to preferred hand) had more explanatory power than dot-probe position (i.e., left or right side of touchscreen), which may have masked any effect of hemispheric lateralization in our study. Therefore, we coded each trial to reflect the location of the dot-probe in relation to the preferred hand of the individual being tested (i.e., position preferred hand: ipsilateral or contralateral; Table 1). The post hoc analysis consisted of an information theoretic approach to evaluate the goodness-of-fit, Akaike Information Criterion (AICc) scores, differences in AICc scores (ΔAICc), and Akaike weights (scales models relative to one another) of the 2 models of interest (one for each stimuli duration; [80, 81]). Specifically, we compared the full model including position preferred hand (replacing the variable dot-probe position) to the original full model for each stimuli duration using the “aictab” function of the “AICcmodavg” package (version 2.2.2; [82]), which ranks the models based on the selected Akaike information criteria.

Results

All 8 monkeys reached the 80% performance training criterion to be considered to participate in the dot-probe experiment. Seven monkeys participated in all 5 sessions that the dot-probe task was administered; monkey B refused to participate in the first session following anesthesia (A + 1d), but participated in the other 4 sessions. Each monkey was able to finish the dot-probe task for each session she participated in (exposed to 144 whole face and scrambled trials per session), except in 1 instance during A + 1d where it was necessary to stop the task and return the monkey to the home cage due to the cage being required for unrelated veterinary purposes (monkey E during A + 1d, exposed to 119 whole face and scrambled trials). Additionally, monkey E’s third test session following prolonged anesthesia occurred after 8 days and was categorized the A + 7d test session. Monkey A’s first test session immediately after experiencing prolonged anesthesia occurred 2 days after and was grouped with the A + 1d data. Mean reaction time across all testing days to the dot-probe replacing whole face and scrambled stimuli for 100 ms trials was 668 ± 252 ms and 709 ± 263 ms for 1,000 ms trials (see online suppl. Table 6 for more detailed information).

Q1: Do We Find Attention Bias during Baseline? – Attention Bias toward Threatening Content at 100 ms (but Not 1,000 ms)

Our first question addressed whether the expected pattern of attention bias toward threat was evident in our study sample at baseline. If so, we predicted faster responses to probes replacing aggressive (congruent) versus neutral (incongruent) stimuli.

The model comparison between the full and null model for 100 ms stimuli duration was significant (LRT: $\chi^2 = 16.41$, $df = 6$, $p = 0.012$). The final model did not include any interactions between congruency, trial type, and dot-probe position, but did indicate significant main effects of each of these variables (Table 2; raw data for each predictor are plotted per monkey in the online suppl. material section “Supplementary Figures of Raw Data”). As predicted, monkeys were faster to respond to congruent trials than incongruent trials (LRT: $\chi^2 = 4.04$, $df = 1$, $p = 0.045$; Table 2; Fig. 2) demonstrating an attention bias toward stimuli with threatening content at baseline. A significant effect of trial type indicated monkeys responded slower to dot-probes replacing whole face stimuli as compared to scrambled stimuli (LRT: $\chi^2 = 7.39$, $df = 1$, $p = 0.007$; Table 2; Fig. 2). However, congruency and trial type did not interact significantly, indicating the overall pattern of attention bias toward threatening content did not differ between the whole face and scrambled stimuli. The final model had a fixed effect variance (marginal R^2) of 0.29, a fixed and random effects variance (conditional R^2) of 0.53, and repeatability measurement of 0.18. Predictor variable power was 90.7% for trial type (confidence interval range: 88.7%–92.4%) and 58.1% for congruency (confidence interval range: 55.0%–61.2%). The model comparison between the full and null model for 1,000 ms stimuli duration trials was nonsignificant (LRT: $\chi^2 = 3.56$, $df = 7$, $p = 0.829$).

Q2: Do We See Affect Mediated Changes in Attention Bias following Prolonged Anesthesia? – Attention Bias Switches to Avoidance of Threatening Content

Our second question focused on whether attention bias changed following prolonged anesthesia. If so, we predicted an interaction between test session (baseline vs. A + 1d) and congruency.

The model comparison between the full and null model for 100 ms stimuli duration trials was significant (LRT: $\chi^2 = 13.95$, $df = 5$, $p = 0.012$). The final model included interactions between test session and congruency (LRT: $\chi^2 = 3.461$, $df = 1$, $p = 0.063$; Table 2), and test session and trial type (LRT: $\chi^2 = 4.643$, $df = 1$, $p = 0.031$; Table 2). On the day immediately following anesthesia, monkeys were slower to respond to congruent trials than incongruent trials, with the opposite pattern seen at baseline (Fig. 3; raw data are plotted by test session per monkey in the online suppl. material section “Supplementary Figures of Raw Data”). This result suggests prolonged anesthesia triggered a change in how monkeys responded to threatening information following prolonged anesthesia com-

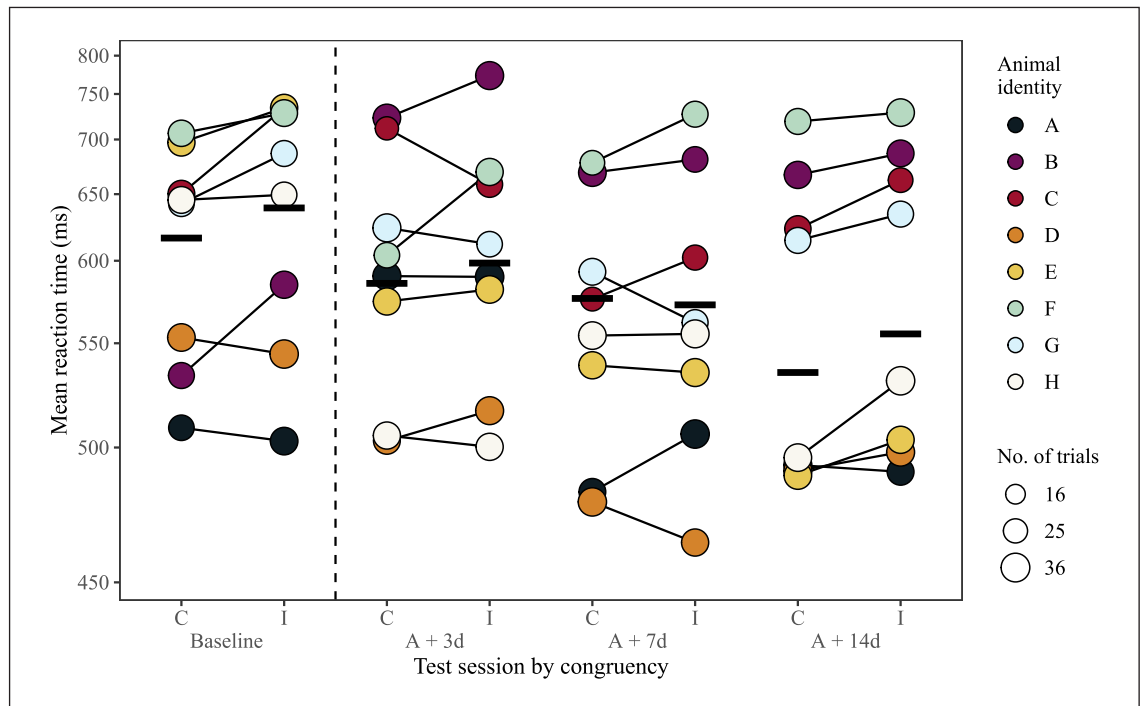


Fig. 4. Investigating when monkeys' responses to congruent and incongruent trials returned to baseline levels following prolonged anesthesia for 100 ms trials. Data shown were taken at baseline, 3- (A + 3d), 7- (A + 7d), and 14-days (A + 14d) after anesthesia. The dashed vertical line separates baseline data from data collected in the 2 weeks following anesthesia. Data taken on the one immediately following anesthesia (A + 1d) are not shown as they were not included in the model. Mean reaction time per monkey to congruent

and incongruent trials per test session, connected by a thin black line. The final model indicates the test sessions occurring 3-, 7-, and 14-days after anesthesia did not differ in their congruency pattern from baseline. The point area indicates the number of trials per condition, ranging from 23 to 35 trials. The Y-axis are scaled according to the transformed data. Model estimates are indicated by thick horizontal lines for each condition when all other predictors are at their mean (either dummy coded or z-transformed).

if where the dot-probe appeared in relation to the monkeys' preferred hand (i.e., position preferred hand) explained the data better than the dot-probe position (left or right on touchscreen) for both stimuli durations. The post hoc information-theoretic analysis for the 100 ms stimuli duration indicated the best model was the original model including dot-probe position over the one including the variable position preferred hand, with an Akaike weight of 0.95 (of 1.00 of the Akaike model weights combined). This analysis provides evidence that the output of the original model including dot-probe position is warranted for stimulus pairs presented for 100 ms (online suppl. Table 4; [80]). In contrast, the model including position preferred hand for the 1,000 ms stimuli duration was the best model in comparison to the original model including dot-probe position, with an Akaike weight of 0.99, indicating evidence as the best model of the two (online suppl. Table 4; [80]).

Discussion

Validating techniques for assessing animal welfare is essential for determining which tools are the most informative, sensitive, and reliable. Such endeavors will help researchers focus on welfare indices that are most useful and thereby enhance the scientific outcomes of projects involving animal research models [83]. Well-established affect-mediated attention bias methods from human cognitive psychology research show promise for use in other animals. Before being used to assess welfare, these tasks must be tested using a context known to change other indices of welfare (e.g., physiological responses). Therefore, we assessed the potential of the dot-probe task for detecting psychological changes, as measured by attention biases, after prolonged anesthesia in adult female long-tailed macaques.

Similar to other dot-probe studies in humans (e.g., [33–36]), we found that the monkeys, when not stressed,

Chapter 5

Comprehensive search filters for retrieving publications on nonhuman primates for literature reviews (filterNHP)

Lauren C. Cassidy^{a,b}, Cathalijn H. C. Leenaars^{c,d,e}, Alan V. Rincon^f, and Dana Pfefferle^{a,b}

^aWelfare and Cognition Group, Cognitive Neuroscience Laboratory, German Primate Center - Leibniz Institute for Primate Research, Göttingen, Germany ^bLeibniz-ScienceCampus Primate Cognition, German Primate Center & University of Göttingen, Göttingen, Germany; ^cInstitute for Laboratory Animal Science, Hannover Medical School, Hannover, Germany ^dDepartment of Animals in Science and Society, Faculty of Veterinary Sciences, Utrecht University, Utrecht, The Netherlands.; ^eSYstematic Review Center for Laboratory animal Research, Department for Health Evidence (section HTA), Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands; ^fDepartment of Psychology, University of Portsmouth, Portsmouth, United Kingdom

This article was published (online: 31 May 2021) in *American Journal of Primatology*, 83(7): e23287 (2021). DOI: 10.1002/ajp.23287.

Contribution to the field

Developing comprehensive search strategies for literature reviews is time-consuming and detail-oriented. These issues can be reduced by the use of thoughtfully, expert-developed search filters on a particular topic. The following study presents comprehensive search filters developed to detect studies involving non-human primates that were highly sensitive to publications referencing non-human primate terminology. Additionally, we created filterNHP, an open-access web-based application, to make these search filters and those for the taxonomic levels of non-human primates easily accessible to other researchers. Use of our comprehensive non-human primate search filters will enhance the quality of non-human primate literature reviews, help refine non-human primate welfare by minimizing the duplication of invasive research, and reduce the time necessary to develop search strategies for future reviews.

Author contributions

Lauren Cassidy, Cathalijn Leenaars, Alan Rincon, and Dana Pfefferle contributed to the conception and design of the study. Lauren Cassidy, Cathalijn Leenaars, and Dana Pfefferle developed the methodology. Lauren Cassidy and Alan Rincon curated the data. Alan Rincon programmed the web-based application, with visualization input from Lauren Cassidy and Dana Pfefferle. Lauren Cassidy performed the analysis and validation, with input from Cathalijn Leenaars and Dana Pfefferle. Lauren Cassidy, Cathalijn Leenaars, and Dana Pfefferle interpreted the data. Lauren Cassidy wrote the manuscript, with revision by all others. All authors read, revised, and approved the submitted version of the manuscript.

Acknowledgments

The authors would like to thank Stefanie Heiduck, an information specialist at the German Primate Center library (Goettingen, Germany; DPZ), for reviewing the search terms included in the NHP filters. Many thanks to Eckhard Heymann, Dietmar Zinner, and Matthias Markolf for beta-testing the functionality of the filterNHP R web-based application and providing insightful comments on the content of the compiled search filters. They would also like to thank Hendrik Eichenauer in the Information Technology department at the DPZ for setting up and maintaining the server to host the filterNHP application and for debugging help.

Statement of ethics

The present study was conducted without the use of NHPs and complies with the American Society of Primatologists Principles for the Ethical Treatment of NHPs.

Conflict of interest statement

This study was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interests.

Funding sources

This study was supported by the German Research Foundation (<http://www.dfg.de>) Research unit 2591 “Severity assessment in animal-based research” (grant numbers: BL953/11-2, GA1475/6-1/2, PF659/5-2, TR447/5-1/2) and the Federal State of Lower Saxony (R2N).

Supplementary material and data

The web application developed in this article can be found here: <https://filternhp.dpz.eu>. The data supporting this article can be found in the online repository, *Göttingen Research Online*. DOI: 10.25625/UTT4SN.

Comprehensive search filters for retrieving publications on nonhuman primates for literature reviews (filterNHP)

Lauren C. Cassidy^{1,2}  | Cathalijn H. C. Leenaars^{3,4,5}  | Alan V. Rincon⁶  | Dana Pfefferle^{1,2} 

¹Welfare and Cognition Group, Cognitive Neuroscience Laboratory, German Primate Center-Leibniz Institute for Primate Research, Goettingen, Germany

²Leibniz-Science Campus Primate Cognition, German Primate Center, University of Goettingen, Goettingen, Germany

³Institute for Laboratory Animal Science, Hannover Medical School, Hannover, Germany

⁴Department of Animals in Science and Society, Faculty of Veterinary Sciences, Utrecht University, Utrecht, The Netherlands

⁵Systematic Review Center for Laboratory animal Research, Department for Health Evidence (section HTA), Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands

⁶Department of Psychology, University of Portsmouth, Portsmouth, UK

Correspondence

Lauren C. Cassidy, German Primate Center, Kellnerweg 4, Goettingen 37077, Germany.
Email: lcassidy@dpz.eu

Funding information

German Research Foundation; Research unit 2591 "Severity assessment in animal-based research", Grant/Award Numbers: BL 953/11-2, GA 1475/6-1/2, PF 659/5-2, TR 447/5-1/2; Federal State of Lower Saxony, Grant/Award Number: R2N

Abstract

Nonhuman primates (NHPs) are widely studied across many scientific disciplines using a variety of techniques in diverse environments. Due to the wide scope of NHP research, substantial overlap in research topics and questions can occur, whose resulting scientific evidence is synthesized by literature reviews. Identifying all relevant research on a particular topic involving NHPs can be difficult and time consuming. By adopting objective search development techniques from systematic reviews, we developed search filters to detect all scientific publications involving NHPs in PubMed, PsycINFO (via EBSCOhost), and Web of Science. We compared the performance of our comprehensive NHP search filters to search strings typical of a novice database user (i.e., NHP simple search strings) and validated their sensitivity by combining these searches with a topic search of cortisol related studies. For all comparisons, our comprehensive NHP search filters retrieved considerably more scientific publications than the NHP simple search strings. Importantly, our comprehensive NHP search filters are easy to use (text can be copied and pasted into the database search engine) and detect the most recent publications that have yet to be indexed by the bibliographic databases queried. Additionally, we developed filterNHP, an R package and web-based application (<https://filterNHP.dpz.eu>), for researchers interested in literature searches involving a taxonomic sub-group of NHPs. filterNHP alleviates time necessary for adapting our comprehensive NHP search filters for a particular NHP sub-group by automating the creation of these search filters. Altogether, our comprehensive NHP search filters and those for taxonomic sub-groups generated by filterNHP will enable swift and easy retrieval of the available scientific literature involving NHPs, and thereby help enhance the quality of new NHP literature reviews that guide future scientific research (new experiments) and public policy (e.g., on welfare and conservation).

KEYWORDS

literature review, nonhuman primates, search filter, systematic review

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *American Journal of Primatology* Published by Wiley Periodicals LLC

1 | INTRODUCTION

Research involving nonhuman primates (NHPs) spans a variety of scientific fields ranging from conservation biology and ecology to cognition and neuroscience. Not only do NHP species fascinate us, but they play a vital role in understanding our own evolution, the natural environment that surrounds us and its biodiversity, and advancing basic (i.e., fundamental) and biomedical research (Estrada et al., 2017; Phillips et al., 2014; Roelfsema & Treue, 2014). While research involving NHPs is often conducted in different environments (e.g., in the wild vs. captivity) using different approaches (e.g., observational vs. experimental), the vast volume of literature can result in substantial overlap in the topics and questions that are investigated and thereby inform the research of other fields. As the literature within these fields is ever-expanding, reliable and objective methods for summarizing evidence across these fields are needed to inform conservation efforts and policy, captive welfare practices, biomedical research, basic research questions, and more involving NHPs.

Generally, literature reviews aim to synthesize scientific evidence across disciplines and provide an overview of this synthesis, while identifying knowledge gaps and informing the direction of new studies. Simple literature searches often begin by identifying and reviewing key sources to the topic being studied, including specific journals, conference proceedings, or high-quality publications (Chapman et al., 2010). Relevant studies often cite additional relevant publications that are subsequently investigated, a method commonly referred to as “snowballing” (Greenhalgh & Peacock, 2005). While simple literature searches and snowballing will identify many relevant articles, publications that use non-standard terminology and are infrequently or never cited will not be found, which may cause bias in the review results. To prevent bias, comprehensive search strategies detecting all relevant publications are needed.

Developing a comprehensive search strategy is time intensive because it involves thoughtful testing of different topic-related terms to evaluate the overall relevance of their search results to the research question (Hausner et al., 2012). Time spent on search development can be saved by using search filters from other reviews or those developed for a specific purpose. For example, search filters have been developed to retrieve studies involving animal research models from PubMed and Embase (de Vries et al., 2011, 2014; Hooijmans et al., 2010), retrieving more studies than the limits provided by the bibliographic databases themselves (e.g., *Limit: Animals* in PubMed; de Vries et al., 2011, 2014; Hooijmans et al., 2010).

In general, standardized search filters are comprised of two parts. The first part consists of topic-relevant standardized search terms (i.e., index terms) from the thesaurus of each online scientific bibliographic source (i.e., databases or platforms hosting multiple databases). The United States National Library of Medicine, for instance, developed MeSH terms (i.e., Medical Subject Headers) for medicine and public health databases (mainly PubMed) that their staff tag (i.e., index) to publications that are deemed as topic-relevant (Bramer et al., 2018). Searching for all primate related studies identified by PubMed, for example, can be executed by using the MeSH term “*Primates*.” Similarly,

PsycINFO, a database of psychological science literature digitalized and indexed by the American Psychological Association (APA), hosts the APA Thesaurus of Psychological Index Terms that allows quick identification of relevant publications. These databases organize index terms into a hierarchical structure by subject category so that more specific terms (i.e., narrow) are nested under more general terms (i.e., broad). For instance, in PubMed the term “*Macaca*” is nested under “*Primates*” in the MeSH term hierarchy. Interestingly, Hooijmans et al. (2010) note that many scientists do not use index terms, potentially due to lack of awareness or confusion surrounding the operation of bibliographic thesauruses, differences in syntax or search guidelines between bibliographic sources, or no universal thesaurus.

The second part of standardized search filters consists of topic-relevant terms to be searched in the title, abstract, and author-defined key words (i.e., TIAB terms). TIAB terms are essential to comprehensive search filters for multiple reasons: some publications may not have been indexed yet, relevant index terms may not be available (e.g., the family *Daubentonidae* is currently not available in the PubMed MeSH database), and indexing may not be completely accurate (e.g., “*Lemur*” is not nested under the category “*Primates (Nonhuman)*” in the APA thesaurus; de Vries et al., 2011; Hooijmans et al., 2010). The TIAB term part of the filter should include all possible spellings and synonyms of the terms of interest to be comprehensive. Specific to NHPs, this means including the scientific name and the common name, spelling variants (e.g., spaces, dashes, and abbreviations), singular and plural, and other synonyms used to reference a species. Covering all possibilities can be quite time consuming and end up in extensive search filters that can be difficult to revise and edit. The string for the TIAB term part for orangutans in PubMed, for example, would be: “*pongidae*[tiab] OR “*ponginae*[tiab] OR “*pongid*[tiab] OR “*pongids*[tiab] OR “*pongo*[tiab] OR “*orangutan*[tiab] OR “*orangutans*[tiab] OR “*orang utan*[tiab] OR “*orang utans*[tiab] OR “*orang-utan*[tiab] OR “*orang-utans*[tiab] OR “*orangutang*[tiab] OR “*orangutangs*[tiab] OR “*orang utang*[tiab] OR “*orang utangs*[tiab] OR “*orangutangs*[tiab].

Multiple scientific bibliographic sources should be searched for relevant publications for literature reviews to be truly comprehensive. A strategy for studies on NHPs may combine results from PubMed, PsycINFO, and the multi-database platform Web of Science for example. While there is substantial overlap in the references found within these sources, this combination would retrieve studies from the medical, neuroscience, (field) biology, psychology, and veterinary fields.

NHP researchers are often interested in focusing on a taxonomic sub-group within the order Primates to answer comparative research questions (e.g., within a taxonomic family). Current search filters including all animal research model species are available as one paragraph of text in PDF format (de Vries et al., 2011, 2014; Hooijmans et al., 2010) which must be tediously searched for relevant terms if search filters on taxonomic sub-groups are desired. Furthermore, while comprehensive search filters provide an excellent reference for creating search strings for a particular sub-group of species (e.g., taxonomic family), sub-group specific terminology is not present to minimize redundant text (e.g., including the term “*old world monkey*” is redundant due to the inclusion of the term

"*monkey[s]*"). This terminology, however, is necessary to include in sub-group search filters as generic terms are too broad (e.g., a search filter for Catarrhini species must include the term "*old world monkey[s]*" instead of "*monkey[s]*" as this term is also used to refer to Platyrrhini species). Also, each bibliographic source has its quirks in how index and TIAB terms are queried (i.e., syntax), adding another layer of complexity for researchers to keep track of. Therefore, solutions for flexibly creating search filters for sub-groups of NHP species and different bibliographic sources are needed.

To aid future literature reviews on NHPs, we developed comprehensive search filters for all NHP species for PubMed, PsycINFO (via EBSCOhost), and Web of Science. Comprehensive NHP search filters were validated by comparing the number of publications they retrieved to search strings that a user with limited search experience (i.e., novice) would use (i.e., NHP simple search strings). Additionally, we combined our comprehensive NHP search filters with a topic related search (i.e., cortisol related studies) and compared search results to those retrieved by the combination of the NHP simple search string and a topic search string.

We provide our comprehensive NHP search filters in Table 1. These comprehensive NHP search filters are also accessible through our R package filterNHP and a web-based application linked to that R package (<https://filterNHP.dpz.eu>). filterNHP can be used to flexibly create comprehensive and NHP taxa specific search filters for three scientific bibliographic sources (PubMed, PsycINFO, Web of Science).

2 | DESCRIPTION

2.1 | Ethics statement

The present study was conducted without the use of NHPs and complies with the American Society of Primatologists Principles for the Ethical Treatment of NHPs.

2.2 | NHP search filter creation and validation

2.2.1 | Search term selection

The comprehensive NHP search filters consist of NHP relevant index terms from the thesaurus of each online bibliographic source (i.e., index term part) and NHP relevant terms to be searched in the title, abstract, and author-defined keywords (if applicable) of bibliographic source publications (i.e., TIAB term part). Scientists with extensive experience working with NHPs, both in the field and in captivity, collated term synonyms and spellings for the search terms relevant to NHPs.

Index term parts were created for PubMed and PsycINFO by identifying relevant index terms from each database's thesaurus (PubMed: MeSH terms; PsycINFO: APA Thesaurus of Psychological Index Terms). Our complete index term list for PubMed included MeSH terms beginning with the category "*Primates*" and included all sub-categories except those relevant to humans (i.e., "*Humans*", "*Neanderthals*"; Figure 1). For PsycINFO, our index term list included all APA index terms in the category "*Primates (Nonhuman)*" and "*Lemur*," where the latter category was not nested under the former (Figure 1). We did not create an index term list for Web of Science as this platform does not have an associated thesaurus.

We identified terms to include in the TIAB term part by referencing the taxonomy and general terminology (e.g., common names) listed in Perelman et al. (2011), Estrada et al. (2017), the taxonomic webpage of the National Center for Biotechnology Information (NCBI; United States National Library of Medicine) website (<https://www.ncbi.nlm.nih.gov/taxonomy>), based off of our own experience, and through discussions with other NHP researchers. Taxonomic nomenclature relevant to extant NHP species were included (i.e., order, suborder, infraorder, parvorder, superfamily, family, subfamily, tribe, genus, subgenus). In cases when a genus name retrieved publications unrelated to NHPs (i.e., "*Mirza*", "*Mico*", "*Nasalis*", "*Carlito*"), we specified the full scientific name (genus and species). We also

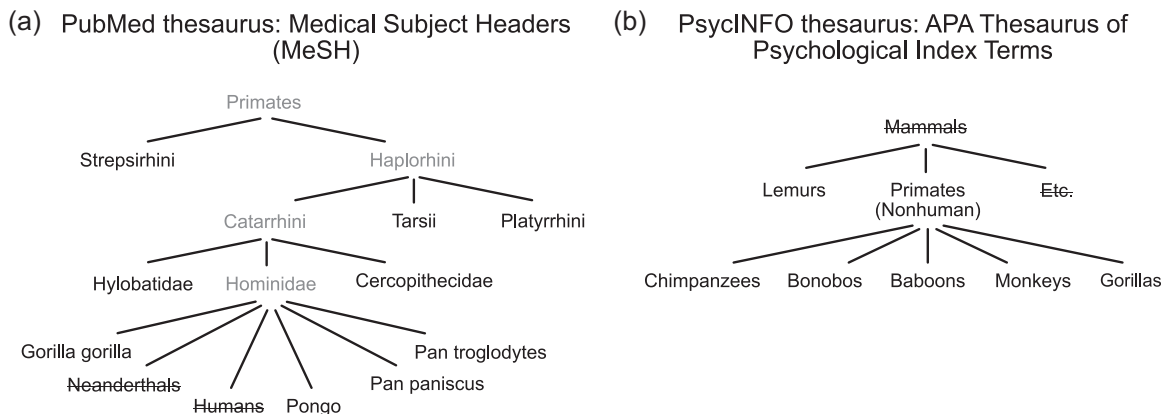


FIGURE 1 Index term hierarchy for NHPs (nonhuman primates) in (a) PubMed and (b) PsycINFO. (a) Medical subject headings (MeSH terms) used by PubMed are structured in a hierarchy where more specific (i.e., narrow) terms are nested under general (i.e., broad) terms. Terms in black are exploded and those in gray are not exploded in the PubMed NHP search filter. Terms with strikethrough text (i.e., "*Humans*", "*Neanderthals*") were not included in the PubMed comprehensive NHP search filter. (b) Hierarchy of the index terms falling underneath "*Mammals*" in the APA Thesaurus of Psychological Index Terms used by PsycINFO. All terms specified were included in the PsycINFO comprehensive NHP search filter except for strikethrough text (i.e., "*Mammals*", "*Etc.*")

included the common names typically used to refer to general groups of NHPs (e.g., “monkey[s]”, “gibbon[s]”). Frequently referenced, but no longer taxonomically correct terminology (e.g., “prosimian[s]”, “callitrichidae”) were also included to ensure that older publications containing these terms are retrieved in searches. It must be noted that search terms can occur in publications irrelevant to NHP species. For example, searching for the term “kipunji” results not only in the retrieval of publications on the NHP species *Rungwecebus kipunji* (including the outdated scientific name of *Lophocebus kipunji*), but also includes one about *Tropostreptus kipunji*, a Tanzanian millipede. These irrelevant publications, however, can be filtered during the search result screening phase. Additionally, we omitted taxonomic and scientific nomenclature relevant to humans (“Humanoid[s]”, “Homo”). TIAB terms were spelled in Latin, American English, British English, and with or without dashes where applicable. Nouns were listed in their singular and plurals forms.

2.2.2 | Search filter syntax

All search terms were combined with the Boolean operator “OR” and surrounded with quotations to detect that exact phrase, which is mainly important for compound terms. To reduce the text of the search filters, we included an asterisk following terms that could be truncated to indicate that the searched word(s) could be followed by any group of characters, including no character. For example, the term “galagid*” detects publications referencing “galagid”, “galagids”, and “galagidae.” We tested truncated terms in all databases to ensure that NHP related publications were retrieved. Below we describe additional syntax specific for each scientific bibliographic source.

PubMed syntax

MeSH index terms in PubMed can be specified to automatically include nested sub-categories by “explosion.” Alternatively, all sub-categories can be omitted by specifying “no explosion.” We specified exploded and unexploded index terms by adding “[mh]” or “[mh:noexp],” respectively. As our search filter targets NHP species and omits humans, the index terms “Primates” and “Catarrhini,” for example, were not exploded, as the undesired index terms “humans” and “neanderthals” fall underneath these categories (note that while using a “NOT human” search may seem an efficient strategy, this would exclude comparative studies involving humans and relevant species; e.g., Kret et al., 2018; Table 1). In contrast, the index term “Platyrrhini” was exploded (Table 1). Search terms for the TIAB part were indicated by adding “[tiab]” to the term, which retrieves publications with this term in the title and/or abstract.

PsycINFO via EBSCOhost syntax

We tailored the syntax for search filters generated for the PsycINFO database for use on the online platform EBSCOhost. Relevant index terms were specified in EBSCOhost by the field “Subjects [exact]” using the modifier “DE” before a list of index terms in parentheses (e.g., Table 1). Please note that while index term explosion is possible

in EBSCOhost by checking the “explode” box, this function only explodes an index term to include sub-categories nested directly (i.e., one sublevel), but not indirectly (i.e., more than one sublevel) underneath. TIAB search terms were listed in parentheses following the modifier “TX” to search the “All Text” field in the PsycINFO database (e.g., Table 1).

Web of science syntax

The Web of Science search filter consists only of search terms from the TIAB term list as this platform does not have an associated thesaurus. We listed search terms in parentheses following the modifier “TS=” (e.g., Table 1). In addition to searching the titles, abstracts, and author-defined keywords, the field “Topic” also searches for terms using the platform’s algorithm called “Keywords Plus.” “Keywords Plus” retrieves publications that are broadly related, but lack the original search terms (e.g., detects publications with terms of other mammal species but no NHP terms), potentially increasing the time necessary for researchers to filtering search results for directly topic-relevant publications.

2.3 | filterNHP: An R package and web-based application for creating NHP search filters

To aid researchers interested in a specific sub-group of NHPs, we created filterNHP, an R package, and interactive web-based application built in R (version 4.0.2) using the Shiny package (version 1.5.0; Chang et al., 2020). The main purpose of filterNHP is to generate comprehensive search filters for NHP taxa (Table 2), a useful tool for anyone interested in conducting a literature search for studies involving a sub-group of NHPs. Additionally, filterNHP allows for search filter terminology to be regularly updated as more relevant NHP terms may be produced or/and identified (e.g., new taxonomic nomenclature). Currently, filterNHP can create search filters for PubMed, PsycINFO via EBSCOhost, and Web of Science. The filterNHP R package can be installed from GitHub by typing `remotes::install_github("avrincon/filterNHP")` into R (note: `remotes` may require prior installation). Additionally, filterNHP is available as a web-based application at: <https://filterNHP.dpz.eu>.

To create search filters for specific sub-groups of NHPs generated by filterNHP, we expanded the TIAB term list to include terms that would have otherwise been redundant to include in the search filters for all NHP species. Specifically, we included common names down to the genus level (e.g., “titi monkey[s]”) and more specific groups of NHPs (e.g., “African monkey[s]”, “Asian monkey[s]”) for search filters where the general term may be omitted because it is too broad (e.g., the term “monkey[s]” is too broad if only African monkeys are of interest). We also added the full scientific name (genus and species) of a species when the genus was also a general term for a taxonomic level (e.g., “lemur”, “colobus”). For genera that were reclassified, we included the full outdated scientific names when identified in the taxonomic browser of the NCBI website or personal communication with other NHP researchers. For example, NHPs from the genus *Leontocebus* were previously classified under the

TABLE 1 Comprehensive NHP search filters for (a) PubMed, (b) PsycINFO (via EBSCOhost), and (c) Web of Science.

a: Comprehensive NHP search filter for PubMed

“catarrhini”[mh:noexp] OR “cercopithecidae”[mh] OR “gorilla gorilla”[mh] OR “haplorhini”[mh:noexp] OR “hominidae”[mh:noexp] OR “hylobatidae”[mh] OR “pan paniscus”[mh] OR “pan troglodytes”[mh] OR “platyrrhini”[mh] OR “pongo”[mh] OR “primates”[mh:noexp] OR “strepsirhini”[mh] OR “tarsii”[mh] OR “allenopithecus”[tiab] OR “allocebus”[tiab] OR “alouatta”[tiab] OR “alouattinae”[tiab] OR “angwantibo”[tiab] OR “anthropoid”[tiab] OR “anthropoidea”[tiab] OR “anthropoids”[tiab] OR “aotes”[tiab] OR “aotidae”[tiab] OR “aotinae”[tiab] OR “aotus”[tiab] OR “ape”[tiab] OR “apes”[tiab] OR “arctocebus”[tiab] OR “ateles”[tiab] OR “atelidae”[tiab] OR “atelines”[tiab] OR “avahi”[tiab] OR “aye-aye”[tiab] OR “baboon”[tiab] OR “baboons”[tiab] OR “bonobo”[tiab] OR “bonobos”[tiab] OR “brachyteles”[tiab] OR “bushbabies”[tiab] OR “bushbaby”[tiab] OR “cacajao”[tiab] OR “callibella”[tiab] OR “callicebinae”[tiab] OR “callicebus”[tiab] OR “callimico”[tiab] OR “callitrichid”[tiab] OR “callitrichinae”[tiab] OR “callitrix”[tiab] OR “callitrichid”[tiab] OR “callitrichidae”[tiab] OR “callitrichide”[tiab] OR “callitrichids”[tiab] OR “callitrichinae”[tiab] OR “capuchin”[tiab] OR “capuchins”[tiab] OR “carlito syrichta”[tiab] OR “catarrhine”[tiab] OR “catarrhini”[tiab] OR “catarrhina”[tiab] OR “catarrhine”[tiab] OR “catarrhini”[tiab] OR “cebid”[tiab] OR “cebidae”[tiab] OR “cebids”[tiab] OR “cebiniae”[tiab] OR “ceboidea”[tiab] OR “cebuella”[tiab] OR “cebus”[tiab] OR “cephalopachus”[tiab] OR “cercocebus”[tiab] OR “cercopithecid”[tiab] OR “cercopithecinae”[tiab] OR “cercopithecine”[tiab] OR “cercopithecini”[tiab] OR “cercopithecoid”[tiab] OR “cercopithecoidea”[tiab] OR “cercopithecoids”[tiab] OR “cercopithecus”[tiab] OR “cheirogaleidae”[tiab] OR “cheirogaleus”[tiab] OR “cheracebus”[tiab] OR “chimp”[tiab] OR “chimpanzee”[tiab] OR “chimpanzees”[tiab] OR “chimps”[tiab] OR “chiromyiformes”[tiab] OR “chiropotes”[tiab] OR “chlorocebus”[tiab] OR “colobidae”[tiab] OR “colobinae”[tiab] OR “colobine”[tiab] OR “colobini”[tiab] OR “colobus”[tiab] OR “cynomolgus”[tiab] OR “daubentonia”[tiab] OR “daubentoniidae”[tiab] OR “doug”[tiab] OR “doucs”[tiab] OR “erythrocebus”[tiab] OR “eulemur”[tiab] OR “euoticus”[tiab] OR “euprimate”[tiab] OR “galagid”[tiab] OR “galago”[tiab] OR “galagoides”[tiab] OR “galagonidae”[tiab] OR “galagos”[tiab] OR “gelada”[tiab] OR “geladas”[tiab] OR “gibbon”[tiab] OR “gibbons”[tiab] OR “gorilla”[tiab] OR “gorillas”[tiab] OR “grivet”[tiab] OR “grivets”[tiab] OR “guenon”[tiab] OR “guereza”[tiab] OR “hapalemur”[tiab] OR “haplorhine”[tiab] OR “haplorhini”[tiab] OR “haplorrhine”[tiab] OR “haplorrhini”[tiab] OR “hominid”[tiab] OR “hominin”[tiab] OR “homininae”[tiab] OR “hominine”[tiab] OR “hominines”[tiab] OR “hominini”[tiab] OR “hominins”[tiab] OR “homoidea”[tiab] OR “hoolock”[tiab] OR “howler”[tiab] OR “hylobates”[tiab] OR “hylobatidae”[tiab] OR “indri”[tiab] OR “indridae”[tiab] OR “indriid”[tiab] OR “indris”[tiab] OR “kipunji”[tiab] OR “lagothrix”[tiab] OR “langur”[tiab] OR “langurs”[tiab] OR “lemur”[tiab] OR “lemurid”[tiab] OR “lemuriform”[tiab] OR “lemuriformes”[tiab] OR “lemuriforms”[tiab] OR “lemurinae”[tiab] OR “lemuroidea”[tiab] OR “lemurs”[tiab] OR “leontideus”[tiab] OR “leontocebus”[tiab] OR “leontopithecus”[tiab] OR “lepilemur”[tiab] OR “lepilemurid”[tiab] OR “lesula”[tiab] OR “lophocebus”[tiab] OR “loriform”[tiab] OR “loriformes”[tiab] OR “lorinae”[tiab] OR “loris”[tiab] OR “lorises”[tiab] OR “lorisid”[tiab] OR “lorisiform”[tiab] OR “lorisinae”[tiab] OR “lorisoid”[tiab] OR “lutung”[tiab] OR “lutungs”[tiab] OR “macaca”[tiab] OR “macaque’s”[tiab] OR “macaque”[tiab] OR “macaques”[tiab] OR “malbrouck”[tiab] OR “mandrill”[tiab] OR “mandrills”[tiab] OR “mandrillus”[tiab] OR “mangabey”[tiab] OR “marmoset”[tiab] OR “marmosets”[tiab] OR “mico argentatus”[tiab] OR “mico chrysoleucus”[tiab] OR “mico emiliae”[tiab] OR “mico humilis”[tiab] OR “mico marcai”[tiab] OR “mico melanurus”[tiab] OR “mico rondoni”[tiab] OR “microcebus”[tiab] OR “miopithecus”[tiab] OR “mirza coquereli”[tiab] OR “mirza zaza”[tiab] OR “monkey”[tiab] OR “monkeys”[tiab] OR “muriqui”[tiab] OR “nasalis larvatus”[tiab] OR “nomascus”[tiab] OR “nycticebus”[tiab] OR “oedipomidas”[tiab] OR “orang utan”[tiab] OR “orang-utan”[tiab] OR “orangutan”[tiab] OR “oreonax”[tiab] OR “otolemur”[tiab] OR “pan paniscus”[tiab] OR “pan troglodytes”[tiab] OR “panin”[tiab] OR “panina”[tiab] OR “panins”[tiab] OR “papio”[tiab] OR “papionini”[tiab] OR “paragalago”[tiab] OR “perodicticinae”[tiab] OR “perodicticus”[tiab] OR “phaner”[tiab] OR “piliocolobus”[tiab] OR “pithecia”[tiab] OR “pithecidae”[tiab] OR “pitheciid”[tiab] OR “pitheciinae”[tiab] OR “pithecinae”[tiab] OR “platyrrhine”[tiab] OR “platyrrhini”[tiab] OR “platyrrhina”[tiab] OR “platyrrhine”[tiab] OR “platyrrhini”[tiab] OR “pleurocebus”[tiab] OR “pongid”[tiab] OR “ponginae”[tiab] OR “pongo”[tiab] OR “potto”[tiab] OR “pottos”[tiab] OR “presbytini”[tiab] OR “presbytis”[tiab] OR “primate”[tiab] OR “primates”[tiab] OR “procolobus”[tiab] OR “prolemur”[tiab] OR “propithecus”[tiab] OR “prosimian”[tiab] OR “prosimii”[tiab] OR “pseudopotto”[tiab] OR “pygathrix”[tiab] OR “rhinopithecus”[tiab] OR “rungwecebus”[tiab] OR “saguinus”[tiab] OR “saimiri”[tiab] OR “saimiriinae”[tiab] OR “sapajus”[tiab] OR “sciurocheirus”[tiab] OR “semnopithecus”[tiab] OR “siamang”[tiab] OR “siamangs”[tiab] OR “sifaka”[tiab] OR “sifakas”[tiab] OR “simians”[tiab] OR “simias”[tiab] OR “simiiform”[tiab] OR “strepsir”[tiab] OR “surili”[tiab] OR “sympalangus”[tiab] OR “talapoin”[tiab] OR “tamarin”[tiab] OR “tamarins”[tiab] OR “tamarinus”[tiab] OR “tarsier”[tiab] OR “tarsiens”[tiab] OR “tarsiid”[tiab] OR “tarsiiform”[tiab] OR “tarsius”[tiab] OR “theropithecus”[tiab] OR “trachypithecus”[tiab] OR “uacari”[tiab] OR “uakari”[tiab] OR “uakaris”[tiab] OR “varecia”[tiab] OR “vervet”[tiab]

b: Comprehensive NHP search filter for PsycINFO (via EBSCOhost)

DE(“baboons” OR “bonobos” OR “chimpanzees” OR “gorillas” OR “lemurs” OR “monkeys” OR “primates (nonhuman)”) OR TX(“allenopithecus” OR “allocebus” OR “alouatta” OR “alouattinae” OR “angwantibo” OR “anthropoid” OR “anthropoidea” OR “anthropoids” OR “aotes” OR “aotidae” OR “aotinae” OR “aotus” OR “ape” OR “apes” OR “arctocebus” OR “ateles” OR “atelidae” OR “atelines” OR “avahi” OR “aye-aye” OR “baboon” OR “baboons” OR “bonobo” OR “bonobos” OR “brachyteles” OR “bushbabies” OR “bushbaby” OR “cacajao” OR “callibella” OR “callicebinae” OR “callicebus” OR “callimico” OR “callitrichid” OR “callitrichinae” OR “callitrix” OR “callitrichid” OR “callitrichidae” OR “callitrichide” OR “callitrichids” OR “callitrichinae” OR “capuchin” OR “capuchins” OR “carlito syrichta” OR “catarrhine” OR “catarrhini” OR “catarrhina” OR “catarrhine” OR “catarrhini” OR “cebid” OR “cebidae” OR “cebids” OR “cebiniae” OR “ceboidea” OR “cebuella” OR “cebus” OR “cephalopachus” OR “cercocebus” OR “cercopithecid” OR “cercopithecinae” OR “cercopithecine” OR “cercopithecini” OR “cercopithecoid” OR “cercopithecoidea” OR “cercopithecoids” OR “cercopithecus” OR “cheirogaleidae” OR “cheirogaleus” OR “cheracebus” OR “chimp” OR “chimpanzee” OR “chimpanzees” OR “chimps” OR “chiromyiformes” OR “chiropotes” OR “chlorocebus” OR “colobidae” OR “colobinae” OR “colobine” OR “colobini” OR “colobus” OR “cynomolgus” OR “daubentonia” OR “daubentoniidae” OR “doug” OR “doucs” OR “erythrocebus” OR “eulemur” OR “euoticus” OR “euprimate” OR “galagid” OR “galago” OR “galagoides” OR “galagonidae” OR “galagos” OR “gelada” OR “geladas” OR “gibbon” OR “gibbons” OR “gorilla” OR “gorillas” OR “grivet” OR “grivets” OR “guenon” OR “guereza” OR “hapalemur” OR “haplorhine” OR “haplorrhine” OR “haplorrhini” OR “hominid” OR “hominin” OR “homininae” OR “hominine” OR “hominines” OR “hominini” OR “hominins” OR “homoidea” OR “hoolock” OR “howler” OR “hylobates” OR “hylobatidae” OR “indri” OR “indridae” OR “indriid” OR “indris” OR “kipunji” OR “lagothrix” OR “langur” OR “langurs” OR “lemur” OR “lemurid” OR “lemuriform” OR “lemuriformes” OR “lemuriforms” OR “lemurinae” OR “lemuroidea” OR “lemurs” OR “leontideus” OR “leontocebus” OR “leontopithecus” OR “lepilemur” OR “lepilemurid” OR “lesula” OR “lophocebus” OR “loriform” OR

(Continues)

TABLE 1 (Continued)

b: Comprehensive NHP search filter for PsycINFO (via EBSCOhost)
<p>"loriformes" OR "lorinae" OR "loris" OR "lorises" OR "lorisid*" OR "lorisiform*" OR "lorisinae" OR "lorisoid*" OR "lutung" OR "lutungs" OR "macaca" OR "macaque's" OR "macaque" OR "macaques" OR "malbrouck*" OR "mandrill" OR "mandrills" OR "mandrillus" OR "mangabey*" OR "marmoset" OR "marmosets" OR "mico argentatus" OR "mico chrysoleucos" OR "mico emiliae" OR "mico humilis" OR "mico marcai" OR "mico melanurus" OR "mico nigriceps" OR "mico rondoni" OR "microcebus" OR "miopithecus" OR "mirza coquereli" OR "mirza zaza" OR "monkey" OR "monkeys" OR "muriqui*" OR "nasalis larvatus" OR "nomascus" OR "nycticebus" OR "oedipomidas" OR "orang utan*" OR "orang-utan*" OR "orangutan*" OR "oreonax" OR "otolemur" OR "pan paniscus" OR "pan troglodytes" OR "panin" OR "panina" OR "panins" OR "papio" OR "papiioni" OR "paragalago" OR "perodicticinae" OR "perodicticus" OR "phaner" OR "piliocolobus" OR "pithecia" OR "pithecidae" OR "pitheciid*" OR "pitheciinae" OR "pithecinae" OR "platyrhine*" OR "platyrhini" OR "platyrrhina" OR "platyrrhine*" OR "platyrrhini" OR "plecturocebus" OR "pongid*" OR "ponginae" OR "pongo" OR "potto" OR "pottos" OR "presbytini" OR "presbytis" OR "primate" OR "primates" OR "procolobus" OR "prolemur" OR "propithecus" OR "prosimian*" OR "prosimii" OR "pseudopotto" OR "pygathrix" OR "rhinopithecus" OR "rungwecebus" OR "saguinus" OR "saimiri" OR "saimiriinae" OR "sapajus" OR "sciurocheirus" OR "semnopithecus" OR "siamang" OR "siamangs" OR "sifaka" OR "sifakas" OR "simians" OR "simias" OR "simiiform*" OR "strepsir*" OR "surili*" OR "symphalangus" OR "talapoin*" OR "tamarin" OR "tamarins" OR "tamarinus" OR "tarsidae" OR "tarsier" OR "tarsiers" OR "tarsiid*" OR "tarsiiform*" OR "tarsius" OR "theropithecus" OR "trachypithecus" OR "uacari*" OR "uakari" OR "uakaris" OR "varecia" OR "vervet*")</p>
c: Comprehensive NHP search filter for Web of Science
<p>TS=("allenopithecus" OR "alloecebus" OR "alouatta" OR "alouattinae" OR "angwantibo*" OR "anthropoid" OR "anthropoidea" OR "anthropoids" OR "aotes" OR "aotidae" OR "aotinae" OR "aotus" OR "ape" OR "apes" OR "arctocebus" OR "ateles" OR "atelidae" OR "atelineae" OR "avahi" OR "aye-aye*" OR "baboon" OR "baboons" OR "bonobo" OR "bonobos" OR "brachyteles" OR "bushbabies" OR "bushbaby" OR "cacajao" OR "callibella" OR "callicebinae" OR "callicebus" OR "callimico" OR "callitrichid*" OR "callitrichinae" OR "callitrix" OR "callitrichid" OR "callitrichidae" OR "callitrichide" OR "callitrichids" OR "callitrichinae" OR "capuchin" OR "capuchins" OR "carlo syrichta" OR "catarhine*" OR "catarrhini" OR "catarrhina" OR "catarrhine*" OR "catarrhini" OR "cebid" OR "cebidae" OR "cebids" OR "cebiniae" OR "ceboidea" OR "cebuella" OR "cebus" OR "cephalopachus" OR "cercocebus" OR "cercopithecid*" OR "cercopithecinae" OR "cercopithecine*" OR "cercopithecini" OR "cercopithecoid" OR "cercopithecoidea" OR "cercopithecoidea" OR "cercopithecoids" OR "cercopithecus" OR "cheirogaleidae" OR "cheirogaleus" OR "cheracebus" OR "chimp" OR "chimpanzee" OR "chimpanzees" OR "chimps" OR "chiromyiformes" OR "chiropotes" OR "chlorocebus" OR "colobidae" OR "colobinae" OR "colobine*" OR "colobini" OR "colobus" OR "cynomolgus" OR "daubentonia" OR "daubentoniidae" OR "douc" OR "doucs" OR "erythrocebus" OR "eulemur" OR "euoticus" OR "euprimate*" OR "galagid*" OR "galago" OR "galagoes" OR "galagoides" OR "galagonidae" OR "galagos" OR "gelada" OR "geladas" OR "gibbon" OR "gibbons" OR "gorilla" OR "gorillas" OR "grivet" OR "grivets" OR "guenon*" OR "guereza*" OR "hapalemur" OR "haplorhine*" OR "haplorhini" OR "haplorrhine*" OR "haplorrhini" OR "hominid*" OR "hominin" OR "homininae" OR "hominine" OR "hominines" OR "hominini" OR "hominins" OR "homoidea" OR "hoolock" OR "howler*" OR "hylobates" OR "hylobatidae" OR "indri" OR "indridae" OR "indriid*" OR "indris" OR "kipunji*" OR "lagothrix" OR "langur" OR "langurs" OR "lemur" OR "lemurid*" OR "lemuriform" OR "lemuriformes" OR "lemuriforms" OR "lemurinae" OR "lemuroidea" OR "lemurs" OR "leontideus" OR "leontocebus" OR "leontopithecus" OR "lepilemur" OR "lepilemurid*" OR "lesula*" OR "lophocebus" OR "loriform" OR "loriformes" OR "lorinae" OR "loris" OR "lorises" OR "lorisid*" OR "lorisiform*" OR "lorisinae" OR "lorisoid*" OR "lutung" OR "lutungs" OR "macaca" OR "macaque's" OR "macaque" OR "macaques" OR "malbrouck*" OR "mandrill" OR "mandrills" OR "mandrillus" OR "mangabey*" OR "marmoset" OR "marmoset" OR "marmosets" OR "mico argentatus" OR "mico chrysoleucos" OR "mico emiliae" OR "mico humilis" OR "mico marcai" OR "mico melanurus" OR "mico nigriceps" OR "mico rondoni" OR "microcebus" OR "miopithecus" OR "mirza coquereli" OR "mirza zaza" OR "monkey" OR "monkeys" OR "muriqui*" OR "nasalis larvatus" OR "nomascus" OR "nycticebus" OR "oedipomidas" OR "orang utan*" OR "orang-utan*" OR "orangutan*" OR "oreonax" OR "otolemur" OR "pan paniscus" OR "pan troglodytes" OR "panin" OR "panina" OR "panins" OR "papio" OR "papiioni" OR "paragalago" OR "perodicticinae" OR "perodicticus" OR "phaner" OR "piliocolobus" OR "pithecia" OR "pithecidae" OR "pitheciid*" OR "pitheciinae" OR "pithecinae" OR "platyrhine*" OR "platyrhini" OR "platyrrhina" OR "platyrrhine*" OR "platyrrhini" OR "plecturocebus" OR "pongid*" OR "ponginae" OR "pongo" OR "potto" OR "pottos" OR "presbytini" OR "presbytis" OR "primate" OR "primates" OR "procolobus" OR "prolemur" OR "propithecus" OR "prosimian*" OR "prosimii" OR "pseudopotto" OR "pygathrix" OR "rhinopithecus" OR "rungwecebus" OR "saguinus" OR "saimiri" OR "saimiriinae" OR "sapajus" OR "sciurocheirus" OR "semnopithecus" OR "siamang" OR "siamangs" OR "sifaka" OR "sifakas" OR "simians" OR "simias" OR "simiiform*" OR "strepsir*" OR "surili*" OR "symphalangus" OR "talapoin*" OR "tamarin" OR "tamarins" OR "tamarinus" OR "tarsidae" OR "tarsier" OR "tarsiers" OR "tarsiid*" OR "tarsiiform*" OR "tarsius" OR "theropithecus" OR "trachypithecus" OR "uacari*" OR "uakari" OR "uakaris" OR "varecia" OR "vervet*")</p>

Note: Index terms are in bold. An asterisk indicates that the search includes words stemming from the truncated word, followed by any group of or no characters. For an explanation of PubMed syntax, refer to the PubMed User Guide section on Search Field descriptions and tags by the National Library of Medicine [<https://pubmed.ncbi.nlm.nih.gov/help/#search-tags>]. For an explanation of EBSCOhost syntax, refer to the EBSCOhost syntax search tips page [https://connect.ebsco.com/s/article/EBSCOhost-API-Health-Library-Query-Syntax-Search-Tips?language=en_US]. For an explanation of Web of Science syntax, refer to the Web of Science Core Collection Help pages describing search operators and rules [<https://images.wobofknowledge.com/WOKRS533JR18/help/WOS/contents.html>].

Abbreviation: NHP, nonhuman primate.

genus *Saguinus*, so a search filter for *Leontocebus* species also specifies their outdated scientific names (e.g., "*Saguinus fuscicollis nigrifrons*" was the previous scientific name for "*Leontocebus nigrifrons*") to detect older publications.

For the creation of NHP sub-group search filters (Table 2) via filterNHP, we specified which index and TIAB search terms should be included, avoiding redundancy when a general term could replace a specific term. A search filter for the infraorder Lemuriformes, for

TABLE 2 The primate order phylogenetic tree.

Suborder	Infraorder	Parvorder	Superfamily	Family	Subfamily	Tribe	Genus	
Strepsirrhini	Lemuriformes			Lemuridae			Lemur, Eulemur, Varecia, Hapalemur, Prolemur	
				Lepilemuridae			Lepilemur	
				Cheirogaleidae			Cheirogaleus, Microcebus, Mirza, Allocebus, Phaner	
				Indridae			Indri, Avahi, Propithecus	
				Daubentoniidae			Daubentonia	
				Lorisiidae	Perodicticinae		Arctocebus, Perodicticus	
					Lorisiinae		Loris, Nycticebus	
					Galagidae		Galago, Euoticus, Galagoidea, Otolemur, Paragalago, Sciuurocheirus	
					Tarsiidae		Carlito, Cephalopachus, Tarsius	
					Cebidae	Callitrichinae	Cebuella, Callibella, Mico, Callithrix, Callimico, Saguinus, Leontocebus, Leontopithecus	
Haplorhini	Tarsiiformes	Platyrrhini		Cebinae			Cebus, Sapajus	
				Saimiriinae			Saimiri	
				Aotinae			Aotus	
				Alouattinae			Alouatta	
				Atelinae			Ateles, Brachyteles, Lagothrix	
				Pitheciinae			Cacajao, Chiropotes, Pithecia	
					Callicebinae		Plecturocebus, Callicebus, Cheracebus	
					Cercopithecoidea	Cercopithecoidea	Cercopithecoidea	Allenopithecus, Miopithecus, Erythrocebus, Chlorocebus, Cercopithecus
						Papionini	Macaca, Cercocebus, Lophocebus, Rungwecebus, Papio, Theropithecus, Mandrillus	
						Colobini	Colobus, Ptilocolobus, Procolobus	
	Hominoidea				Presbytini		Semnopithecus, Simias, Trachypithecus, Presbytis, Pygathrix, Rhinopithecus, Nasalis	
							Hoolock, Hylobates, Nomascus, Symphalangus	
					Ponginae		Pongo	
					Hominiinae	Gorillini	Gorilla	
						Hominini	Pan, Homo	

Note: Referencing Perelman et al. (2011). The genus *Homo* was not included as a search term in the comprehensive NHP search filters or as category in filterNHP.

example, would include the broad term “*lemur[s]*”, but not “*Lemur catta*.” However, the former term in this example would be omitted and the latter term included in a search filter for the family *Lemuridae* due to species in other taxonomic families being referred to as lemurs.

Currently, filterNHP automatizes the creation of NHP taxa search filters for three bibliographic sources (PubMed, PsycINFO, and Web of Science) using their specific syntax and index terms where relevant. Using the function `filter_nhp()` in R, users can specify a bibliographic source (argument: `source`) and the broadest taxonomic level(s) of interest (argument: `taxa`) as a vector to receive a complete search filter in the console that can be easily copied and pasted to the search engine of the corresponding bibliographic source. Spelling of the specified bibliographic source and taxa must follow that written in the R documentation and Table 2. Users can also choose to omit certain taxa if desired by listing these taxa as a vector (argument: `omit`). For example, researchers investigating species in the infraorder Haplorhini and omitting the superfamily Hominoidea in Web of Science, would indicate “Web of Science” in the `source` argument, “Haplorhini” in the `taxa` argument, and “Hominoidea” in the `omit` argument (Figure 2). In this example (`filter_nhp(source = "Web of Science," taxa = "Haplorhini," omit = "Hominoidea")`), the generated search filter still includes broad terms, such as Simiiformes and Catarrhini (which encompasses Hominoidea). This behavior of the `omit` argument is useful because while it may return more *irrelevant* publications, it may also return more *relevant* publications that would have gone undetected, hence, improving search strategy comprehensiveness. If the user wishes for a search filter omitting broader terms, however, then the more specific taxonomic branches of interest should be listed as a vector following the `taxa` argument (e.g., `taxa = c("Cercopithecoidea", "Platyrrhini", "Tarsiiformes")`).

The web-based Shiny application is an interactive, user-friendly implementation of the package filterNHP that can be used without installing R. The workflow and all functionality described for running the function in R is also possible in the application (Figure 2) and is described on the webpage (<https://filterNHP.dpz.eu>).

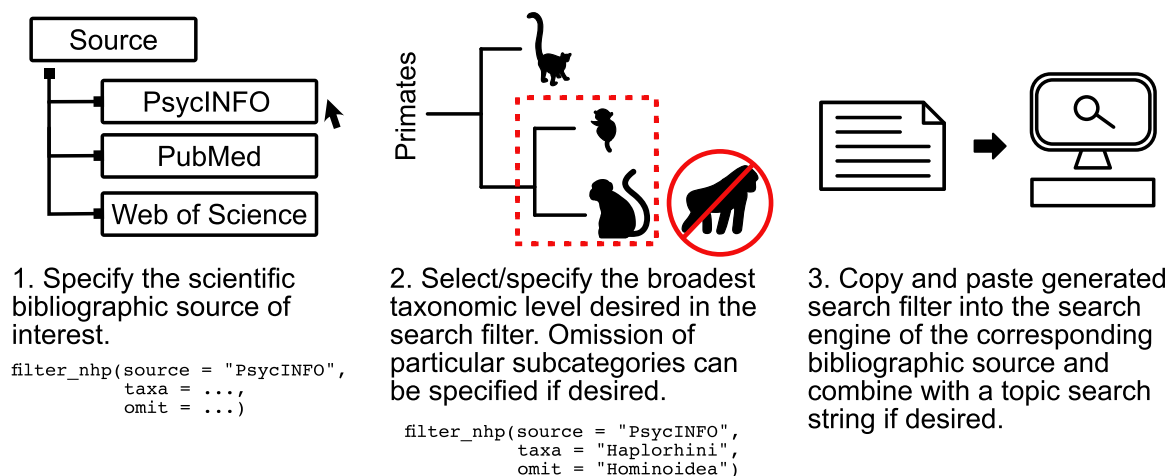


FIGURE 2 Visual representation of the filterNHP workflow

3 | EXAMPLE

3.1 | Implementation of comprehensive NHP search filters and NHP simple search strings for comparison

We conducted searches using our comprehensive NHP search filters for three bibliographic sources (PubMed, PsycINFO via EBSCOhost, Web of Science; Table 1) and compared their results with those of the NHP simple search strings (i.e., string of search terms typical of a novice to literature searches; Table 3). Initial searches were conducted for all comprehensive NHP search filters and NHP simple search strings. For Web of Science, we used the search box of the *Advanced Search* page as field modifiers were already specified for the search filters and simple search strings. Subsequent searches were combined using Boolean operators. Table 4 summarizes the total number of results per search.

3.2 | Performance evaluation and validation of the NHP search filters

Following the initial searches, we checked how much overlap in the search results occurred between our comprehensive NHP search filters and the NHP simple search strings. All publications detected by the NHP simple search strings were also detected by our comprehensive NHP search filters (search #3; Table 4). Moreover, across bibliographic sources, our comprehensive NHP search filters detected 1.6 to 3.3 times more publications than the NHP simple search strings (search #4; Table 4).

We also tested the utility of our comprehensive NHP search filters by performing a topic search of cortisol related studies to highlight the role of the search filters in developing comprehensive literature reviews on topics involving NHPs. Therefore, the NHP simple search strings and our comprehensive NHP search filters were combined with the topic search string (containing cortisol specific terminology) using the Boolean operator “AND” (see Table 3 for topic search strings). We compared the performance of the topic

search with the NHP simple search strings and our comprehensive NHP search filters for search result overlap and to see how many more publications the latter search retrieved over the former search (Table 4). All publications detected by the topic search using the NHP simple search strings were also detected by the topic search with our comprehensive NHP search filters (search #7; Table 4). The topic search yielded 1.4 to 2.3 times more publications when combined with our comprehensive NHP search filters than when combined with the NHP simple search strings (search #8; Table 4).

4 | COMPARISON AND CRITIQUE

Finding all the relevant research involving NHPs is not a trivial task for those interested in conducting literature reviews as the NHP scientific literature is massive and widespread. Providing a complete and unbiased overview of a scientific topic, must begin with clearly defined, comprehensive search strategies. Search filters developed for a taxonomic group can increase the retrieval of potentially relevant publications within a bibliographic source and decrease the time needed for search development (e.g., de Vries et al., 2011, 2014; Hooijmans et al., 2010).

Within this article, we present search filters for detecting potentially relevant publications involving extant NHP species in PubMed, PsycINFO via EBSCOhost, and Web of Science. Search term lists forming the basis of our comprehensive NHP search filters were created by consulting the scientific literature, the NCBI taxonomy website, based off our own expertise, and conversations with other NHP researchers. As a result of this detail-oriented investigation, we anticipate our NHP search filters to be comprehensive.

We have made our NHP search filters available as text (Table 1 of this article) and through an R package and web-based application (filterNHP) so that they are widely accessible to researchers. However, note that, modifications to our NHP search filters are needed to search other databases and platforms. Adaptations to other bibliographic sources are a future development goal of the filterNHP R package and web-based application. In the meantime, researchers using other sources are advised to contact an information specialist for help.

To test our filters, we compared their performance to NHP simple search strings, alone and in combination with a topic search on cortisol, in PubMed, PsycINFO, and Web of Science. We detected 1.4 to 3.3 times more publications with our filters than the simple

TABLE 4 Results of literature searches by the NHP simple search strings, our NHP search filters, and these in combination with a topic search for PubMed, PsycINFO (via EBSCOhost), and Web of Science

Search	Query	PubMed	PsycINFO (via EBSCOhost)	Web of Science
All NHPs				
#1	NHP simple search string	79,312	35,131	94,474
#2	Comprehensive NHP search filter	249,606	56,107	279,877
#3	#1 AND #2	79,312	35,131	94,474
#4	#2 NOT #1	170,294	20,976	185,403
All NHPs + Topic search string^a				
#5	Topic AND #1	1299	747	1501
#6	Topic AND #2	2856	1038	3008
#7	#6 AND #5	1299	747	1501
#8	#6 NOT #5	1627	291	1507

Note: Search performed on April 12, 2021. AND: records present in both searches; all records retrieved by the simple searches are also retrieved by the new comprehensive search filters; NOT: records present in the first search, but not in the second.

Abbreviation: NHP, nonhuman primate.

^aTopic search strings found in Table 3.

search strings when tested alone and in combination with the cortisol topic search. As publications detected by each search must mention at least one relevant term specified, these large performance differences are driven by the inclusion of a far greater number of relevant index and TIAB terms in our comprehensive NHP search filters than those in the NHP simple search strings. Simple search strings limited to general terms would not detect publications where only the species scientific name is mentioned in title or/and abstract for example. Furthermore, simple search strings may unintentionally overlook alternative spellings, synonyms, or plurals that are easy to forget or miss.

By creating filterNHP as a package available in R and as a web-based application, we sought to improve upon current animal

TABLE 3 NHP simple and topic search strings for PubMed, PsycINFO (via EBSCOhost), and Web of Science for all NHP species

Database	NHP simple search string	Topic search string
PubMed	"Primates" [mh:noexp] OR "nonhuman primates"[tiab] OR "monkeys"[tiab] OR "great apes"[tiab]	hydrocortisone [mh] OR cortiso *[tiab] OR hydrocortiso *[tiab] OR epicortiso *[tiab] OR glucocortico *[tiab]
PsycINFO (via EBSCOhost)	DE("Primates [Nonhuman]" OR "Monkeys") OR TX ("nonhuman primates" OR "monkeys" OR "great apes")	DE(hydrocortisone) OR TX(cortiso * OR hydrocortiso * OR epicortiso * OR glucocortico *)
Web of Science	TS=("nonhuman primates" OR "monkeys" OR "great apes")	TS=(cortiso * OR hydrocortiso * OR epicortiso * OR glucocortico *)

Note: Topic search strings for PubMed and PsycINFO adapted from a systematic review protocol (Smith & Leenaars, 2020). Index terms are in bold. An asterisk indicates that the search includes words stemming from the truncated word, followed by any group of or no characters.

search filters by providing an easy way to parse our comprehensive NHP search filters by sub-groups of NHPs. To obtain such a search filter would otherwise require the researcher to comb through the text of our filters for relevant terms and brainstorm additional terms that may be relevant. While we still recommend the latter process, filterNHP substantially reduces the time needed to revise and edit our comprehensive NHP search filters when required. An additional advantage of filterNHP is that the index term and TIAB term lists that form the basis of generated search filters can be easily updated and adapted as the scientific literature evolves.

Our comprehensive NHP search filters and any from our filterNHP package and web-based application can easily be implemented by electronically copying and pasting a filter into the search box of its corresponding database or platform and clicking *Search*. We recommend that search filters for Web of Science be pasted directly into the search box of the *Advanced Search* page as fields and Boolean operators are already specified. In all bibliographic sources, each new search is assigned a number in the *Search history* page or section. Thereafter, topic searches can be easily combined with the NHP search filter using the operator "AND" to join a string of topic-related search terms and the corresponding search history number (or full filter if desired). This method of combining searches using the search history can also be used to split a large search filter into multiple parts, which may occasionally be necessary when databases or platforms are experiencing heavy use and are slow to retrieve results.

We encourage researchers to review the search terms in the comprehensive NHP search filters and adapt the filters to their own specific needs. Researchers with special interest in extinct NHP taxa, for example, will need to add relevant search terms as we included terminology for extant taxa only. Additionally, filterNHP is set up to produce subgroup specific search filters down to the genus taxonomic level. Therefore, the search term lists that inform the creation of subgroup search filters currently do not include an exhaustive list of species common names, which researchers may want to add if these names do not add redundancy. For example, adding the species common name "*Coppery titi monkey*" to the search filter of the subfamily *Callicebinae* would be redundant as the term "*titi monkey*" is already included, but would be necessary to add for a search filter of the genus *Callicebus*. Finally, we considered adding terms written in languages other than English or Latin (for scientific nomenclature) to the search term lists (Amano et al., 2016), but found that added commonly used phrases (e.g., "*Affe*," meaning "*monkey*" in German) did not retrieve more results. As the databases and platforms use English as their internal language, search results comprise articles in other languages without adding language-specific terms. For example, our Web of Science comprehensive search retrieved articles in 26 non-English languages.

5 | CONCLUSION

Our comprehensive NHP search filters and the filterNHP R package and web-based application will enable researchers to search for NHP related scientific publications swiftly and easily. Furthermore, search filters, such

as ours promote the standardization of literature searches for specific topics (here NHPs) and the use of objective approaches for conducting literature reviews. As the scientific literature on NHPs is constantly expanding across many different disciplines, reliable and objective methods for synthesizing scientific evidence are crucial for determining how robust scientific phenomenon are, improving research reproducibility, reducing animal use and research studies, guiding future research, and informing public policy.

ACKNOWLEDGMENTS

The authors would like to thank Stefanie Heiduck, an information specialist at the German Primate Center library (Goettingen, Germany; DPZ), for reviewing the search terms included in the NHP filters. Many thanks to Eckhard Heymann, Dietmar Zinner, and Matthias Markolf for beta-testing the functionality of the filterNHP R web-based application and providing insightful comments on the content of the compiled search filters. They would also like to thank Hendrik Eichenauer in the Information Technology department at the DPZ for setting up and maintaining the server to host the filterNHP application and for debugging help. This study was supported by the German Research Foundation (<http://www.dfg.de>) Research unit 2591 "Severity assessment in animal-based research" (grant numbers: BL953/11-2, GA1475/6-1/2, PF659/5-2, TR447/5-1/2) and the Federal State of Lower Saxony (R2N).

CONFLICT OF INTERESTS

This study was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interests.

AUTHOR CONTRIBUTIONS

Lauren C Cassidy completed conceptualization (equal); data curation (equal); formal analysis (equal); investigation (lead); methodology (equal); project administration (lead); visualization (equal); writing original draft (lead); writing review and editing (lead). Cathalijn Leenaars completed conceptualization (equal); formal analysis (supporting); methodology (equal); supervision (supporting); validation (equal); writing review & editing (supporting). Alan V Rincon completed conceptualization (equal); data curation (equal); software (Lead); visualization (equal); writing review and editing (supporting). Dana Pfefferle completed conceptualization (equal); formal analysis (supporting); funding acquisition (lead); methodology (equal); supervision (lead); validation (equal); visualization (supporting); writing review and editing (supporting).

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1002/ajp.23287>

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at <https://filterNHP.dpz.eu> and the code at <https://github.com/avrincon/filterNHP>. The findings of this study are openly available

at <https://filterNHP.dpz.eu>. Supporting data can be downloaded at <https://doi.org/10.25625/UTT4SN>. The code for the R package and web-based application can be openly accessed at <https://github.com/avrincon/filterNHP>.

ORCID

Lauren C. Cassidy  <https://orcid.org/0000-0001-8680-0284>

Cathalijn H. C. Leenaars  <https://orcid.org/0000-0002-8212-7632>

Alan V. Rincon  <http://orcid.org/0000-0001-6181-0152>

Dana Pfefferle  <https://orcid.org/0000-0003-4827-0360>

REFERENCES

- Amano, T., González-Varo, J. P., & Sutherland, W. J. (2016). Languages are still a major barrier to global science. *PLOS Biology*, 14(12), e2000933. <https://doi.org/10.1371/journal.pbio.2000933>
- Bramer, W. M., Rethlefsen, M. L., Mast, F., & Kleijnen, J. (2018). Evaluation of a new method for librarian-mediated literature searches for systematic reviews. *Research Synthesis Methods*, 9(4), 510–520. <https://doi.org/10.1002/jrsm.1279>
- Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2020). *Package shiny: Web Application Framework for R. R Package Version 1.5.0*. Retrieved from <https://CRAN.R-project.org/package=shiny>
- Chapman, A. L., Morgan, L. C., & Gartlehner, G. (2010). Semi-automating the manual literature search for systematic reviews increases efficiency. *Health Information & Libraries Journal*, 27(1), 22–27. <https://doi.org/10.1111/j.1471-1842.2009.00865.x>
- de Vries, R. B. M., Hooijmans, C. R., Tillema, A., Leenaars, M., & Ritskes-Hoitinga, M. (2011). A search filter for increasing the retrieval of animal studies in Embase. *Laboratory Animals*, 45(4), 268–270. <https://doi.org/10.1258/la.2011.011056>
- de Vries, R. B. M., Hooijmans, C. R., Tillema, A., Leenaars, M., & Ritskes-Hoitinga, M. (2014). Updated version of the Embase search filter for animal studies. *Laboratory Animals*, 48(1), 88–88. <https://doi.org/10.1177/0023677213494374>
- Estrada, A., Garber, P. A., Rylands, A. B., Roos, C., Fernandez-Duque, E., Di Fiore, A., Nekaris, K. A., Nijman, V., Heymann, E. W., Lambert, J. E., Rovero, F., Barelli, C., Setchell, J. M., Gillespie, T. R., Mittermeier, R. A., Arregoitia, L. V., de Guinea, M., Gouveia, S., Dobrovolski, R., ... Li, B. (2017). Impending extinction crisis of the world's primates: Why primates matter. *Science Advances*, 3(1), e1600946. <https://doi.org/10.1126/sciadv.1600946>
- Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *British Medical Journal*, 331(7524), 1064–1065. <https://doi.org/10.1136/bmj.38636.593461.68>
- Hausner, E., Waffenschmidt, S., Kaiser, T., & Simon, M. (2012). Routine development of objectively derived search strategies. *Systematic Reviews*, 1(1), 19. <https://doi.org/10.1186/2046-4053-1-19>
- Hooijmans, C. R., Tillema, A., Leenaars, M., & Ritskes-Hoitinga, M. (2010). Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. *Laboratory Animals*, 44(3), 170–175. <https://doi.org/10.1258/la.2010.009117>
- Kret, M. E., Muramatsu, A., & Matsuzawa, T. (2018). Emotion processing across and within species: A comparison between humans (*Homo sapiens*) and chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, 132(4), 395–409. <https://doi.org/10.1037/com0000108>
- Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A., Kessing, B., Pontius, J., Roelke, M., Rumpel, Y., Schneider, M. P., Silva, A., O'Brien, S. J., & Pecon-Slattery, J. (2011). A molecular phylogeny of living primates. *PLOS Genetics*, 7(3), e1001342. <https://doi.org/10.1371/journal.pgen.1001342>
- Phillips, K. A., Bales, K. L., Capitanio, J. P., Conley, A., Czoty, P. W., 't Hart, B. A., Hopkins, W. D., Hopkins, W. D., Hu, S. L., Miller, L. A., Nader, M. A., Nathanielsz, P. W., Rogers, J., Shively, C. A., & Voytko, M. L. (2014). Why primate models matter. *American Journal of Primatology*, 76(9), 801–827. <https://doi.org/10.1002/ajp.22281>
- Roelfsema, P. R., & Treue, S. (2014). Basic neuroscience research with nonhuman primates: A small but indispensable component of biomedical research. *Neuron*, 82(6), 1200–1204. <https://doi.org/10.1016/j.neuron.2014.06.003>
- Smith, B., & Leenaars, C. H. C. (2020). *Systematic review of adrenaline, corticosterone and cortisol in microdialysates*. SyRF. Retrieved from <https://drive.google.com/file/d/1MAZrq5oNieNyp9ZkV4hYA29SkM6Emxd/view>

How to cite this article: Cassidy, L. C., Leenaars, C. H. C., Rincon, A. V., & Pfefferle, D. (2021). Comprehensive search filters for retrieving publications on nonhuman primates for literature reviews (filterNHP). *Am J Primatol*, 83, e23287. <https://doi.org/10.1002/ajp.23287>

Chapter 6

General Discussion

The central aim of this thesis was to propose and develop more objective and animal-centric methods of welfare and severity assessment in captive animals. The field of animal welfare science has historically refrained from recognizing the subjective experiences and affective states of animals as informative to animal well-being, but interest in their value is rising. In this last chapter, I first briefly summarize the main findings of the previous chapters. Then I bring together these studies to discuss their significance to their respective fields and the field of animal welfare. Additionally, I identify and discuss some commonalities found across the experimental studies of this thesis (Chapter 2, Chapter 3, Chapter 4). Finally, I propose some future avenues of research for the methods that I developed within this thesis.

6.1 Summary of chapters

In Chapter 2, we proposed and tested the Choice-based Severity Scale (CSS), a novel concept for severity assessment in animals derived from choice-based preference testing methodologies. We conducted a Choice-based Severity Assessment by offering adult male rhesus macaques (*Macaca mulatta*) a choice to perform a basic experimental task between two conditions common in systems neuroscience research: a cage or lab condition. Generally, the monkeys had a strong preference for performing the task in the cage condition, irrespective of the amount of reward per trial provided. However, one individual's preference could be shifted by changing the type of reward, suggesting inter-individual differences in their subjective evaluations of these two conditions. In the CSS test, we found that the amount of reward per trial could be used to convince the monkeys to choose a more difficult task over an easier task. Collectively, these results support the CSS concept as a severity assessment method with the capacity to objectively rank and scale welfare parameters across different domains of animal welfare.

In close association to Chapter 2, Chapter 3 explored whether adult male rhesus macaques are able to learn associations between abstract stimuli and their delayed positive reinforcement. In the delayed reinforcement task of Chapter 3, the monkeys were given a choice between abstract stimuli whose associated reward was delivered up to 10 minutes after the choice. We found that the monkeys did exhibit markers of long-delay learning (i.e., preference for the stimulus delivering higher reward, continued task engagement), even when the abstract stimuli were

novel. Not only do our findings provide further insight into the long-delay learning capabilities of rhesus macaques, but they provide support that the macaques tested in Chapter 2 understood the link between the abstract stimuli and procedures we offered.

Chapter 4 explored whether the dot-probe attention bias task could detect changes in attention bias following prolonged anesthesia in adult female long-tailed macaques (*M. fascicularis*), a context known to modulate attention and trigger physiological arousal in non-human primates (Bethell et al., 2012a; Lee et al., 2010; Novak et al., 2013; Whitten et al., 1998). Within this task, stimulus pairs of threatening and neutral facial expressions of unknown conspecifics and their scrambled counterparts were presented at different durations to capture information about attention allocation and changes in attention bias. We found evidence that the monkeys were more vigilant to threatening stimuli for briefly presented stimulus pairs during a period of putative low arousal (i.e., baseline). The monkeys' attention bias deviated from this baseline pattern (becoming avoidant of threatening stimuli) on the day immediately following anesthesia, but had returned to baseline by the third day following anesthesia. The scrambled stimuli included in our dot-probe task also elicited attention bias effects likely due to retaining the shape of their whole face counterparts, and thus emphasize the importance of careful stimuli preparation and task design. With refinements and further validation, our findings in Chapter 4 indicate that the dot-probe attention bias task can offer insight into the affective states of non-human primates.

In Chapter 5, we took a step beyond directly improving animal welfare and severity assessment methods to develop comprehensive non-human primate search filters that can help refine and reduce non-human primate experimentation through comprehensive literature reviews. Our comprehensive non-human primate search filters were more sensitive to publications referencing terms related to non-human primates than simple search strings (i.e., search strings typical of a person with limited literature search experience). Additionally, we created filterNHP, a web-based application that makes the comprehensive non-human primate search filters easily accessible to other researchers and flexibly generates search filters for other non-human primate taxonomic levels. Outside of helping to refine non-human primate welfare, our search filters can improve the search strategies of literature reviews in other scientific fields involving non-human primates and inspire the development of other taxon-specific search filters.

6.2 Discussion of choice-based preference and attention bias testing

As reviewed in Chapter 1, current animal welfare and severity assessments aim to integrate multiple measures of an animal's physiological and psychological well-being to form an overall picture of its welfare. Presently, these assessments do not account for the animal's perspective and affect. The development of methods sensitive to the subjective experiences and affective states of animals can shed light on which elements of captivity and experimental procedures

they perceive as the most severe. To do so, an animal's choices and/or differences in attention bias between conditions differing by one element can be compared.

6.2.1 Using choice-based preference testing to tap into animals' subjective experiences

The core motivation of the CSS concept developed in Chapter 2 was to reform the development of welfare and severity assessments to reflect the animals' perspective rather than the experimenter's. Currently, these assessments are plagued by argument-by-analogy (Habedank et al., 2018). We cannot directly measure suffering as it is an internal mental state. Therefore, we try to place ourselves "in the other's shoes" to assess their mental state. When a particular reaction is observed, we relate those observations to distinct mental states that we have experienced, and therefore assume that the other is in a comparable state (Juthe, 2005). Unfortunately, argument-by-analogy is often used to interpret behavioral and physiological parameters that are supposed to be objectively measurable (Habedank et al., 2018). For example, two spatial learning tasks (i.e., water maze task: mouse is placed in water and must find a submerged platform; Barnes maze: a circular platform with holes close to the edge, where only one hole leads to the mouse's home cage) applied in mice are classified by local authorities as differing in severity despite both causing significant increases in glucocorticoids (i.e., stress hormones, Habedank et al., 2018). This difference in classification stems from the anthropocentric view of how it would feel to be thrown in a pool of water as one task forces mice to swim whereas the other does not (Habedank et al., 2018). Such anthropocentric judgements are natural but are not scientifically well-founded. Therefore, we must be cautious as these views seep into the animal welfare and severity assessments that we construct.

Do our perspectives of welfare criteria match those of animals? Often there is good empirical evidence behind the ranking of criteria within a welfare parameter and indeed, our sense of what matters may reflect what the animals view as important to their well-being. However, there may be elements of welfare criteria that we are not able to detect. Choice-based preference testing has been recommended as a means to test for preferences between environment-based welfare parameters (e.g., food, enrichment) to shape environmental enrichment from the animal's perspective (Alligood et al., 2017; Fernandez et al., 2004; Fernandez & Timberlake, 2019; Mellen & Sevenich MacPhee, 2001) and improve training outcomes (e.g., Gray et al., 2019). The development of tools like the CSS extends the application of choice-based preference testing by assessing preferences between different procedure-based welfare parameters such as husbandry and experimental procedures. Until we tested the CSS concept in the Choice-based Severity Assessment of Chapter 2, choice-based preference testing had not been applied to procedure-based welfare parameters in the scientific literature as far as I know. Thus, the development of the CSS and its broader application to different types of welfare parameters is a step forward towards improving animal welfare and severity assessment.

To my knowledge, the CSS concept in Chapter 2 is the first severity assessment to propose

a common framework for ranking and scaling welfare parameters across different welfare domains in a more comparative and objective way. Specifically, we proposed to determine the amount of reward necessary to pay animals to choose different welfare criteria. Depending on the animals' choices, the amount of reward is adjusted for each option until the animal perceives the combinations of the options and their respective rewards as equal (i.e., oscillating around a point of subjective equality). This adaptive approach was adopted from human psychophysics experiments to determine perceptual thresholds (Kingdom & Prins, 2010; e.g., Leek, 2001) and forms the basis of automated training protocols to shape complex behaviors in non-human primates (e.g., Berger et al., 2018; Calapai et al., 2022). In the Choice-based Severity Assessment of Chapter 2, we found that all rhesus macaques exhibited a strong preference for performing a task in the cage condition over the lab condition, irrespective of large differences in the amount of reward (potential issues discussed further in the next paragraph). Interestingly, we found that changing the type of reward in the cage condition (from juice to water) did influence one monkey's choices, which suggests individual differences in how the options are evaluated. With these findings, we can assuredly rank the cage condition as having a lower relative severity than the lab condition from the monkeys' perspective, and they hint that aspects of reward can counterbalance the costs of welfare criteria.

So why did the amount of reward not seem to have an effect in the Choice-based Severity Assessment of Chapter 2? In contrast to the Choice-based Severity Assessment, we did find that rhesus macaques select options associated with substantially more reward than others resulting in little reward despite task difficulty (Chapter 2), up to 10 min of delay in the delivery of reward (Chapter 3), and the introduction of *novel* abstract stimuli (Chapter 3). Potentially, the monkeys perceived the amount of reward provided to offset the costs of the conditions as irrelevant because they could engage in (i.e., work) as many trials as they desired within 2 hours. Alternatively, it may have been difficult for the monkeys to detect reward amount changes due to the delay between the choice and receipt of the consequences (approximately 10 minutes). But the validity of this concern is limited as our findings in Chapter 3 indicate that rhesus macaques can learn such associations involving long-delay. Collectively, these findings suggest that scaling options by reward amount is feasible, and that further development of the CSS concept is warranted.

A major strength of the CSS experimental design (i.e., CSS protocol) in Chapter 2 is that it is not limited to offering choices between basic options, but also between more complex options (i.e., compound options). Procedures, such as the lab condition for example, often involve multiple steps that have different associated costs and benefits. Abstract stimuli can be used to represent the whole procedure through positive reinforcement training and can help reduce environmental influences on choice behavior (e.g., physical positioning of devices associated with the procedure). Interestingly, the monkeys were more likely to correctly complete option stimuli training trials for the cage condition (99 % correct on average) than the lab condition (77 % correct on average) during the training phase of the experiment. This finding suggests that the monkeys had a preference for the cage condition due to the first difference in cost

between the two procedures: the degree of movement restraint involved to receive a small fluid reward. Further into testing, the compound options directly reflected the complete in-house procedures of a neuroscience laboratory (performing a task in a cage or lab condition once a day). Such procedures generally consist of multiple steps that could affect the subjective utility of the procedure (e.g., transport, movement restraint, temporary social isolation). For accurate welfare and severity assessment, it is crucial that welfare criteria are assessed under realistic conditions. Thus, the CSS protocol allows for animals to make informed decisions on which an accurate picture of experienced severity (e.g., pain, discomfort, distress) can be built.

The CSS protocol in Chapter 2 reinforces the link between compound options and their corresponding abstract stimuli. Presently, the CSS protocol consists of two reference sessions¹ (one per option) prior to each choice session, resulting in three days per cycle of testing. Reference sessions (one per option), while time-intensive to include, are necessary to choice-based preference testing as they remind animals of the full consequences of compound options and their link to the abstract stimuli used to represent them. Including reference sessions (or trials) is a common practice in cognitive choice tasks (e.g., Morel et al., 2017). Nevertheless, the current CSS protocol could be modified so that multiple choice sessions follow reference sessions, for example.

Across two experiments in Chapter 3, the monkeys discovered and reliably selected a higher value stimulus despite a delay of up to 10 minutes between the selection of a stimulus and the delivery of its associated positive reinforcement. These findings further support that the complications of learning information tied to abstract stimuli under conditions of long-delay (e.g., intervening events, maintain information in memory) are limited in macaques. Furthermore, the monkeys were selective in whether they finished the trial of the chosen stimulus, where highly rewarded were completed more often than low rewarded stimuli. Such selective behavior suggests that the monkeys sustained their commitment to highly rewarded stimuli and retained information about their quality (i.e., understood the stimuli associations with reward, Hayden, 2016). These findings are important to the interpretations of Chapter 2 as they support that the monkeys understood the link between the abstract stimuli and the conditions they represented.

Establishing a predictable chain of events between an abstract stimulus and its delayed consequences is important for enhancing learning in animals (reviewed in Lattal, 2010). In Chapter 2, the procedures for the conditions were conducted in a routine and stereotypic manner, thus linking the selected abstract stimulus to its consequences through a predictable chain of events. Similarly, learning was likely facilitated in Chapter 3 by the features of the “expanding clock” stimulus that reminded the monkeys of their choice and made it possible to track the passage of time. Collectively, the combination of knowledge about the long-delay learning capabilities of animals and clear associations between simple stimuli and complex options open up a lot of possibilities for institutional refinement. For example, behavioral

¹During reference sessions, the monkeys were given a choice between one option stimulus that resulted in either the cage or lab condition, or a stimulus that resulted a short timeout.

measures of welfare in rhesus macaques were improved when husbandry feeding times were signaled by a reliable auditory stimulus that disassociated feeding with out-of-sight caretaker activity (Gottlieb et al., 2013b). Such associative strategies could be applied to procedures involving delay as long as the duration lies within the long-delay learning capabilities of the target animal species.

One of the proposed gold-standards for good animal welfare and quality of life is to provide animals choices to exercise control over their daily life (Leotti et al., 2010; Sambrook & Buchanan-Smith, 1997; Schapiro & Lambeth, 2007). Having the opportunity to choose has been known to improve other individual-based welfare parameters (e.g., physiological: Arce et al., 2010; Behringer et al., 2014; Owen et al., 2005; behavior: Buchanan-Smith & Badihi, 2012; Kurtycz et al., 2014; Owen et al., 2005). There are a couple of considerations to keep in mind, however. First, offering animals choices between procedures and compensating with additional reward is not usually realistic nor compatible with research or/and husbandry protocols in everyday contexts (Habedank et al., 2018). For example, neuroscience experiments often require many trials in a cognitive task to be conducted which would likely be hindered if additional reward is given. Accordingly, the monkeys in Chapter 2 spent much less time working in the lab condition (9 drops of reward per correct trial) on average than in the cage condition (1 drop of reward per correct trial). However, choice-based preference tests like the CSS concept can be applied prior to research involvement to understand the impact these procedures have on animals. Such information is valuable to inform and appropriately construct the ranking and scaling of welfare parameters within welfare and severity assessments. Secondly, animals (and humans) do not necessarily choose what is best for them as subjective utility sometimes deviates from objective utility (Dawkins, 2006; Kahnau et al., 2020). For example, an animal may prefer to eat a sugar cube over a stick of celery, regardless of the possible negative long-term health consequences. Furthermore, preferences also do not necessarily imply that the animal will suffer without access to this preferred resource, particularly if it is a luxury (Dawkins & Beardsley, 1986). Still, these kinds of preferences are informative and can be used to improve motivation and training outcomes (e.g., Gray et al., 2019).

6.2.2 The dot-probe attention bias task is sensitive to changes in affective state

Reasonably, a limitation to the CSS concept that we likely cannot overcome is that we have yet to find a way to offer animals choices between different individual-based welfare parameters (e.g., body weight, alopecia). Specifically, different states of these parameters cannot be experienced simultaneously. Fortunately, affect-sensitive methods can provide insight into psychological well-being where choice-based preference tests cannot. In Chapter 4, we tested the dot-probe attention bias task as tool for detecting affect-mediated changes in attention bias.

Similar to humans (reviewed in Bar-Haim et al., 2007), we found that female long-tailed

macaques were more attendant to threatening stimuli than neutral stimuli when stress was putatively low in Chapter 4. These biases were detected for stimuli presented for the short stimulus duration (100 ms), but not the longer duration (1000 ms). Studies in other non-human primate species have found mixed evidence for detecting attention biases using the dot-probe task (see Appendix C for an overview table). As these studies use a variety of stimuli (e.g., color vs. greyscale, faces vs. whole bodies) and presentation durations, it is difficult to say whether these differences are due to experimental technicalities or are species specific. Given our results, we suggest that attention biases in non-human primates may be best captured using briefly presented color stimuli.

To validate the dot-probe task as a measure of psychological well-being, we tested the monkeys following prolonged anesthesia, a known physiological stressor in non-human primates (Lee et al., 2010; Novak et al., 2013; Whitten et al., 1998). We found that the monkeys' baseline pattern of vigilance to threatening stimuli changed to an avoidance these stimuli on the day immediately following prolonged anesthesia. This change emerged despite typical post-anesthesia side effects (e.g., reaction time slowing, other general motor deficiencies). Such variation could be accounted for as data for all task conditions (i.e., position of the dot-probe relative to the threatening stimulus) were collected within each session. Thus, this change in attention bias reflects changes in the monkeys' psychological well-being. Avoidance to threat is a pattern of attention bias also found in dot-probe studies of humans who experienced chronic stressors such as combat deployment and rocket attack (e.g., Sipos et al., 2014; Wald et al., 2011). The changes in attention biases of our study also follow the similar patterns of attention bias in a looking-time study of male rhesus macaques who also experienced anesthesia (Bethell et al., 2012a). Our findings suggest that the reaction time data from the dot-probe task can capture similar information about overt attention processes (i.e., involving saccades to a stimulus) for shorter stimulus presentation durations when less detailed information about attention is necessary. Broader application of the dot-probe task comparing sets of stimuli in combination with looking-time studies will help elucidate what types of stimuli and stimulus durations capture attention biases best for the species being tested.

Prolonged anesthesia did not have long-lasting effects on the monkeys' psychological well-being as measured by our dot-probe attention bias task. Attention biases had returned to baseline levels (i.e., attendant to threat) by the third day following prolonged anesthesia, a pattern which persisted through the end of testing two weeks later. Such quick recovery may be expected for procedures that occur infrequently and do not generally result in pain (e.g., self-administration of analgesics in mice differs depending on surgery status: Pham et al., 2010). In contrast, recovery from surgical procedures or/and chronic stressors may result in a protracted pattern of threat avoidance depending on the level of postoperative discomfort experienced. For example, the level of combat exposure in military personnel was not only predictive of symptoms of post-traumatic stress disorder and anxiety, but a higher likelihood to avoid threatening stimuli in a dot-probe task (Sipos et al., 2014). If such patterns are typical and can be verified in animals, attention biases could provide much needed insight into

psychological well-being where other welfare measures cannot, especially given that animals often hide obvious signs of suffering (e.g., pain: Landa, 2012).

Differences in the cognitive and affective load of the dot-probe task's stimuli were evident in the monkeys responses to the task. The monkeys were slower to respond to whole face stimuli than scrambled stimuli. This finding is unsurprising given that the processing faces is known to be more cognitively demanding than other objects (e.g., pictures of houses: Holmes et al., 2005). However, the scrambled stimuli still elicited an attention bias effect. While we took care to equate threatening and neutral faces for basic image features (e.g., brightness, contrast), this pattern suggests scrambled stimuli retained affective features of their whole face counterparts. Notably, only one other dot-probe study in non-human primates has included scrambled stimuli (Kret et al., 2018). Our findings emphasize the importance of careful stimuli preparation and the inclusion of control stimuli in the dot-probe task design to ensure effects are driven by the affective content of the stimuli.

Why did the scrambled stimuli also elicit attention bias in Chapter 4? We suspect that face shape, which was retained in the scrambled stimuli, may have captured attention as the threatening stimuli tended to have a higher height-to-width ratio than neutral stimuli. Specifically, threatening images exhibited open-mouth threat expressions of unknown conspecifics (a common aggressive facial display in macaques), which lengthens the face due to the jaw dropping open. As basic shape cues are likely one of the fastest initial cortical image processing steps, those cues in the scrambled stimuli may have allowed for the original facial expression to be estimated ultra-fast given that was the only affective information available to process (Murray et al., 2021). For example, shape cues in line drawings of facial expressions are recognizable by humans (e.g., Etcoff & Magee, 1992) and are features used to categorize facial expressions (e.g., Sormaz et al., 2016). Others have posited that shape information may be one of the most important cues to the perception and recognition of facial expressions (Bruce & Young, 2012; Calder et al., 1996). Importantly, the lack of difference in attention bias by stimulus type in Chapter 4 does not preclude that the dot-probe effect is based on the affective content of stimuli.

6.2.3 Commonalities across the experimental studies

There are several commonalities across the experimental studies of this thesis (Chapter 2, Chapter 3, Chapter 4). The first and foremost is the influence of individuals. The starkest individual differences were found in Chapter 2, where one individual was generally more responsive to changes in reward contingencies across the two experiments (Choice-based Severity Assessment, CSS test) than the other two tested. Additionally, there was some evidence that the two individuals we tested in Chapter 3 had different strategies for how to deal with trials where the low reward option was chosen (i.e., either favoring to abort or leave these trials uncompleted). Changes in attention bias were also stronger for some individuals

than for others in Chapter 4. Such variation is to be expected as individuals react differently to internal and external stimuli and situations (e.g., Coleman, 2012; Howarth et al., 2021; Izzo et al., 2011; Palmer et al., 2022; Sloan Wilson et al., 1994). Furthermore, the findings of Chapter 2 and Chapter 4 highlight that individuals might have different thresholds for coping with the common procedures we tested. Those individuals who require substantially more reward to cope with less desirable procedures or react more strongly in attention bias tasks could be of greatest concern from a welfare perspective. A dot-probe task in surgical patients, for example, found that those most avoidant of pain stimuli prior to operation were more likely to continue to experience high intensity pain in the months following the operation (Lautenbacher et al., 2010). Interestingly, we found in a response slowing study (co-authored over the course of my Ph.D: Bethell et al., 2019) that fearful temperament was predictive of avoidance or behavioral inhibition to touch threatening stimuli. Collectively, these findings suggest that certain individuals may need additional attention from animal care staff to cope with procedures that are less desirable or/and involve putatively heightened stress. Given that animals do not perceive and experience procedures in the same way, methods of welfare and severity assessments should ideally accommodate such differences to reflect an individual's experienced severity more accurately.

The second general commonality across the experimental studies in this thesis is that the monkeys were able to clearly distinguish the visual stimuli that were presented. Specifically, they treated the stimuli differently either by their choices (Chapter 2 and Chapter 3) or differences in their reaction times (Chapter 4). These findings suggest that the monkeys perceived the abstract and facial stimuli as meaningful representations (Fagot et al., 2000; Fagot et al., 2010). Non-human primates are capable of learning to differentiate a variety of categories (e.g., animal or non-animal: Roberts & Mazmanian, 1988; food or nonfood: Fabre-Thorpe et al., 1998; tree or non-tree, fish or non-fish: Vogels, 1999; ordinal numbers: Orlov et al., 2000) and are more attentive towards putatively relevant stimuli in looking-time experiments (e.g., reviewed for faces in Parr, 2011; reviewed for many stimuli categories in Winters et al., 2015). However, what is presently unclear is exactly what stimulus features the monkeys extracted as relevant in these experiments. This question exceeds the scope of the studies in this thesis, but it bears consideration with respects to stimuli preparation for future experiments. Our approach for stimuli preparation in Chapter 2 and Chapter 3 was to make the abstract stimuli distinguishable in multiple ways to facilitate the formation of an association between each stimulus and its respective outcome (Chapter 2: condition-relevant pictures formatted on differently colored squares; Chapter 3: differently shaped and colored stimuli). In contrast, Chapter 4 used biologically relevant stimuli of neutral and aggressive facial expressions to assess changes in attention bias. The intrinsic relevance of such stimuli and their irrelevance to the task (i.e., touching the dot-probe) from the monkeys' perspective are the core strength of the dot-probe attention bias task that allow repeat testing to occur. Overall, careful stimuli selection and preparation in future welfare and severity assessment studies will enhance the validity and reliability of developing methods in animals.

Third, there was a strong supplementary influence of position (e.g., left vs. right, up vs. down) across the experimental studies in this thesis. The choice behavior of rhesus macaques during the pilot experiments in Chapter 2 was strongly influenced by aspects of the environment (overview of pilot experiments in Appendix A). For example, we could not disentangle preferences for the cage-based option from a preference for a quadrant of the testing compartment when the options were positioned vertically as monkeys have a natural tendency to sit in higher locations (Clarence et al., 2006). During the CSS test in Chapter 2 and the delayed positive reinforcement task in Chapter 3, the rhesus macaques had a general preference for selecting the stimulus on the left, which was the dominant hand for all three monkeys. Similarly, the long-tailed macaques were faster to respond dot-probes positioned ipsilateral to their dominant hand in the dot-probe attention bias task in Chapter 4. These commonalities emphasize the importance of counterbalancing stimulus position in cognitive task design as we implemented in our studies and accounted for in our analyses.

The fourth commonality across experiments was that most cognitive tasks (with the exception of the task in lab condition) were administered to the monkeys close to or in their home environment. This testing arrangement took advantage of the existing infrastructure and did not require a separate apparatus to be built which would have involved transporting the monkey. Those monkeys tested in the testing compartment adjacent to their home environment were extensively trained to enter in exchange for a food reward. Thus, the familiarity and proximity to conspecifics of this environment generally minimized arousal (Habedank et al., 2018; Kahnau et al., 2022) and was likely a benefit to the cage condition in Chapter 2. But it is important to note that an animal's subjective utility of an environment (or option) could change. Such influences were temporarily apparent after the monkeys experienced prolonged anesthesia in Chapter 4 as facility limitations necessitated the use of the testing compartment for the administration of the procedure. Specifically, on the day immediately following anesthesia the monkeys were more hesitant to enter the testing compartment and took longer on average to complete the dot-probe attention bias task in comparison to the baseline period (baseline: 20.9 ± 9.1 minutes; the day immediately following anesthesia: 88.7 ± 49.4 minutes). Possibly, the experience temporarily influenced the monkeys willingness to participate in the task. Similarly, bottlenose dolphins (*Tursiops truncatus*) willingness to participate in positive reinforcement training is predictive of changes in health status (Clegg et al., 2019). Chronic social stress in rats (*Rattus norvegicus*) is also associated with reduced locomotor and exploratory activity in behavioral tasks, suggesting a loss of motivation (Rygula et al., 2005). Importantly, the monkeys seemed to have recovered by the third day after anesthesia (taking 29.5 ± 21.0 minutes to complete the dot-probe task), potentially due to repeated exposure to the compartment in association with positive reinforcement. In Chapter 3, the monkeys were tested in their main home cage. This setting possibly helped the monkeys cope with the long delay periods of the task as there were other activities they could carry out (e.g., foraging, observing conspecifics). Subjective experiences may also have more subtle impacts on choice behavior. For example, one of the pilot experiments in Chapter 2 found that the monkeys' choices were substantially influenced by position in the testing compartment,

likely due to one of the quadrants being associated with the cage squeeze (i.e., apparatus used to restrain monkeys during veterinary procedures) despite its rare use. Ultimately, such influences are critical to consider in the study design of cognitive testing in animals as we did in the experimental setup design of Chapter 2 (e.g., by counterbalancing stimuli position).

A fifth common feature to recognize across the experimental studies in this thesis is the small sample size. While not ideal, these numbers are the reality for the basic research contexts that we are working in (i.e., Chapter 2 and Chapter 3: neuroscience; Chapter 4: functional imaging). Moreover, small sample sizes should not prevent us from devising new ways of assessing animal welfare. The methods and findings that I have presented in this thesis provide working proof-of-concepts that lay the foundation for future studies to build on. Additional studies will help improve measures of welfare and severity to build a fuller picture of captive animal psychological well-being.

Lastly, the procedures we tested throughout this thesis (i.e., location to perform a cognitive task, prolonged anesthesia) were multi-componential (i.e., compound resources). In other words, each procedure was comprised of different elements (or steps) that have a greater or lesser impact on psychological well-being. Anesthesia, for example, often involves overnight social group separation to control food intake in addition to the anesthetization and recovery procedure. At this time, we cannot say exactly which elements of these procedures affect psychological well-being the most as these procedures are not usually broken down into smaller elements. Refinement can nevertheless target those elements which likely have the biggest impact, and their success can be evaluated using the methods that I have proposed in this thesis. Providing a social partner prior to and post-anesthesia, for example, may help buffer the effects of stress on psychological well-being if the other individual's well-being is not substantially compromised and the risk of social injury is low (e.g., Gilbert & Baker, 2011; Pham et al., 2010). Overall, the constant goal of animal welfare science is to continually evaluate the impact of and improve upon existing captive management and research practices (Lloyd et al., 2008; Prescott et al., 2017; Rennie & Buchanan-Smith, 2006a; Rennie & Buchanan-Smith, 2006b, 2006c).

6.3 Thoughtfully developed search filters can enhance literature search strategies

The comprehensive search filters for studies involving non-human primates that I developed in Chapter 5 were highly sensitive to topic-relevant studies from PubMed, PsycINFO, and Web of Science. Indeed, the comprehensive non-human primate search filters found 1.4 to 3.3 times more relevant scientific publications than simple search strings when tested alone and in combination with the a topic-relevant string (i.e., terms related to cortisol), respectively. We also combined early versions of the comprehensive non-human primate filters with a topic-relevant string to conduct a qualitative systematic review investigating methods for determining intrinsic preferences in non-human primates. This practical application of

the comprehensive non-human primate search filters resulted in the retrieval of more than 7000 unique scientific publications, suggesting that the search strategy was comprehensive and highly sensitive. Of these publications, 754 have been identified by two independent evaluators as relevant to our study and are currently being extracted for preference test-relevant information. We expect that this systematic review will be a useful resource for others interested in developing choice-based preference tests in animals.

In parallel, we created filterNHP in Chapter 5, a publicly available web-based application that provides easy access to these comprehensive non-human primate search filters. Additionally, filterNHP flexibly generates search filters for the taxonomic levels of the primate order, with the exception of *Homo* species. Researchers can use this literature search tool to reduce the time necessary to develop and implement their own search strategies for future literature reviews. Other automated literature search tools (e.g., ‘litsearchr’ in R) have found that the process of developing a search strategy, conducting the search, and assembling results could be reduced from approximately 17-34 hours to under 2 hours (Grames et al., 2019). Furthermore, use of our comprehensive non-human primate search filters will help avoid typical literature search pitfalls such as selection bias (Haddaway et al., 2015), failing to select a comprehensive set of terms (e.g., including terms with British and American English spellings, singular and plurals, historical taxonomic nomenclature), and lack of bibliographic source-specific knowledge (e.g., use of Medical Subject Headers in PubMed, Salvador-Oliván et al., 2019). Moreover, our comprehensive non-human primate search filters have the potential inspire the development of those for other taxa and to standardize search strategies that will improve the reproducibility, specificity, and objectivity of future literature reviews (Hausner et al., 2012; Stansfield et al., 2017). Importantly, literature reviews using comprehensive search strategies may prevent studies from being duplicated or/and provide useful information (e.g., for power analyses) that could limit the number of animals needed in future studies (Macleod et al., 2005; Pound et al., 2004).

6.4 Future outlook

There are several promising avenues of future research for the methods that I developed for welfare and severity assessment in this thesis. In the next sections, I first touch on several areas where the CSS and affected-mediated attention bias tasks can be expanded to. Then, I elaborate further on several considerations for the future development of the dot-probe attention bias task as an animal welfare and severity assessment measure. As all welfare and severity assessment tools need validation, I propose to combine the methods developed in this thesis with other welfare measures and with each other to verify our findings and provide further insight into psychological well-being experienced during the procedures tested in Chapter 2. Finally, I highlight some promising welfare assessment methods under development in non-human primates.

6.4.1 Expanding the application of the Choice-based Severity Scale and dot-probe attention bias task

Naturally the application of the CSS and dot-probe attention bias task within and across different conditions, settings, and species is needed. Contexts and conditions that foster positive animal well-being and boredom are comparatively understudied in contrast those where negative well-being is experienced, but are important to consider in welfare assessment (positive contexts: Boissy et al., 2007; Crump et al., 2018; Held & Špinka, 2011; Clark, 2011; Meagher, 2019; boredom contexts: Špinka & Wemelsfelder, 2011). The CSS and dot-probe attention bias task can also be applied more broadly to other settings such as zoos and livestock. Arguably these settings may involve less severe procedures, but it is important to recognize that they have other associated stressors that basic and biomedical settings do not (e.g., zoo visitors, resource production). While zoo-based cognitive research has exploded in the last 10 years (Clark, 2017; reviewed in Clegg, 2018; Hopper, 2017; Hopper, 2022), the application of affect-mediated attention bias tasks is lacking and warrants further development and adaptation to other species (reviewed in Crump et al., 2018). The CSS concept can also be adapted to compare environment- and procedure-based welfare parameters in other animal species. Fortunately, captive animals can be conditioned to associate more complex procedures and stimuli with simple, species-relevant stimuli (Habedank et al., 2018). Conditioned place preference tests, for example, have a long history of associating compartments with rewarding and aversive stimuli (reviewed in Tzschentke, 1998, 2007).

6.4.2 Considerations for the future development of the dot-probe attention bias task

Determining the dot-probe attention bias task's repeatability and sensitivity to other welfare parameters that may diminish or promote welfare is needed (Bartlett & Frost, 2008; Bland & Altman, 1986). Some human studies have called into question the test-retest reliability of the dot-probe task (Price et al., 2015; Schmukle, 2005; Staugaard, 2009; Waechter et al., 2014), finding low reliability but consistent attention biases across a couple days of testing (Aday & Carlson, 2019). While these findings are likely due to differences between individuals and affective state (Staugaard, 2009), they suggest that cautious interpretation and further investigation with the dot-probe attention bias task is necessary. The systematic application of the dot-probe attention bias task to other aspects of animal welfare, for example, would help determine the task's sensitivity. If sensitive to other welfare parameters, the task could reveal whether the degree and valence of attention biases (e.g., measured by the difference in reaction times between task conditions; see Chapter 4) can be used as a means to rank and scale welfare components objectively. Attention biases following differently valenced contexts in rhesus macaques suggest that ranking different contexts is at least feasible (Bethell et al., 2012a). Social housing conditions, for example, could be an apt welfare parameter to test with the dot-probe attention bias task as these conditions have known effects on other measures of well-being (e.g., frequency of abnormal behavior exhibited during observation of a veterinary

procedure: Gilbert & Baker, 2011), and can be arranged in a variety of ways (e.g., solitary, pairs, group). Such conditions could also be tested within and across individuals to rigorously test the robustness of the dot-probe task.

Likewise, it is important to determine the generalizability and reproducibility of the dot-probe attention bias task across facilities and species (Howarth et al., 2021; Kilkenney et al., 2014). Although this task may be best suited for non-human primate species, it is feasible to train other animals to use touchscreens (e.g., dogs: Wallis et al., 2017; tortoises: Mueller-Paul et al., 2014) and species-relevant stimuli can be incorporated if the species' visual systems perceive touchscreen images as intended (Egelkamp & Ross, 2019). Alternatively, adaptations or other attention bias tasks that tap into other sensory modalities (e.g., olfactory, gustatory, haptic, auditory) may provide a better means of assessing attention bias (reviewed in Crump et al., 2018). For example, Trevarthen et al. (2019) developed an attention bias task for mice that involved trials where application of a mildly threatening and mildly attractive stimulus were applied as they returned to their home cage. A response slowing paradigm tested in Japanese macaques (*M. fuscata*) found that the monkeys were slower to touch images of conspecific faces on days when noise was elevated due to a loud event in comparison to days where normal levels of noise was experienced (Cronin et al., 2018). Collectively, tasks capturing affect-mediated attention biases are promising tools offering insight into animal psychological well-being (Crump et al., 2018).

6.4.3 Validating the Choice-based Severity Scale and dot-probe attention bias task with other welfare parameters

How do the methods that were developed in this thesis relate to other welfare parameters? In Chapter 4, we relied on a previously verified stressor to validate if our dot-probe task could detect changes in attention bias and hence, psychological well-being. New welfare measures are also often validated by verifying changes in other parameters simultaneously. While our focus in Chapter 2 was to develop the CSS methodology using behavior, it would be possible to investigate other welfare parameters concurrently. Individual-based welfare parameters, such as physiology (e.g., elevated heart rate, elevated cortisol) and behavior (e.g., yawning), could be compared between the cage and lab conditions we tested in Chapter 2. Such parameters could also be applied as animals are conducting the dot-probe attention bias task to see if elevated levels occur in relation to the affective manipulation. Variation in autonomous responses (e.g., pupil diameter: Bradley et al., 2008; Henderson et al., 2018; skin conductance: Gatti et al., 2018; blink rates: Ballesta et al., 2016; skin temperature: Kuraoka & Nakamura, 2011; Froesel et al., 2020; heart rate: Unakafov et al., 2018) would likely provide the most detailed and continuous information about online changes to well-being. Such measures in rhesus macaques are responsive to the presentation of affective auditory and visual stimuli for example (e.g., heart rate: Froesel et al., 2020; skin temperature: Kuraoka & Nakamura, 2011). Not only could such parameters be evaluated across sessions, but within session variation could help determine the usefulness of certain interventions (e.g., receipt of grooming in non-human primate chair: Taira

& Rolls, 1996). Autonomic system measurements require careful selection and interpretation, however, as the degree of movement restriction may influence measures sensitive to arousal and do not discriminate emotional valence (Paul et al., 2005). Importantly, the selection of other welfare parameters to validate the methods that I developed in this thesis should be non-invasive to minimize training and stress, consequently resulting in improved animal welfare (Froesel et al., 2020).

6.4.4 Combining the Choice-based Severity Scale with measures of attention bias

Several interesting questions could be explored by combining the non-human primate protocol with measures of attention bias. After stimuli are equated for basic image features, does the stimulus of the preferred or less-preferred condition capture attention (e.g., gaze) first? Systematic differences in the orientation of reflexive attention would suggest that attentional resources are prioritized to one stimulus, hence condition, over the other (Öhman & Mineka, 2001), analogous to threatening stimuli in the dot-probe attention bias task. Total duration of looking-time could also provide some insight in the role of attention biases in choice as it is often interpreted as a measure of visual preference (reviewed in Winters et al., 2015). For example, human participants more frequently choose stimuli that are experimentally presented for longer than those presented for shorter durations (Shimojo et al., 2003), which suggests that longer durations of looking-time towards a particular stimulus may also elicit similar choice behavior. Interestingly, Wilson et al. (2019) were able to relate looking-time to preferences in a choice task for images of food and objects over landscapes, but not for social images (e.g., kin vs. nonkin, young vs. old) in long-tailed macaques. The usefulness of reference sessions could also be maximized by testing if attention bias as measured by the dot-probe task differs between the conditions that are being tested. A recent study in laying hens (*Gallus domesticus*) used a comparable triangulation approach to test a judgement bias task (i.e., to measures differences in cognitive biases) in combination with preference tests and other candidate welfare parameters (e.g., serum blood glucose levels), finding some consistent alignment between these measures across time (Paul et al., 2022). Applying a similar multi-methodological approach could identify if attention bias does indeed differ between the conditions we tested, which would provide support for the proposal that these conditions do leave traces on the affective lives of animals and their psychological well-being (Lewejohann et al., 2020). Additionally, do patterns of attention bias differ between sessions where an option is electively chosen (i.e., choice sessions) versus those where there is only one viable option (i.e., reference sessions)? Such explorations would offer further insight into the effect of choice on animal welfare.

6.4.5 Other methods of welfare and severity assessment under development in non-human primates

Given the high stakes of non-human primate research, the development of new methods of welfare and severity assessment are crucial to regularly and accurately evaluating the impact of applied research and captive management practices. Consequently, there are several new measures of welfare and severity under development in non-human primates that I would like to highlight. First, facial expressions and behaviors indicative of pain may be useful for determining when pain alleviating interventions should be applied (Descovich, 2017; Descovich et al., 2019). Such indices have been widely developed and applied in other species (e.g., grimace scale: Cohen & Beths, 2020). Second, combining welfare metadata and neuroimaging may offer insight into the effects of different external factors on brain structure (e.g., social network size: Sallet et al., 2011; Testard et al., 2022; hierarchy: Noonan et al., 2014; early adversity: Howell et al., 2019), other individual-specific influences (Poirier et al., 2021), and improving the design chronic implants (Ahmed et al., 2022; Basso et al., 2021). Third, understanding the ‘cumulative severity’ of management and research practices on non-human primates is essential as these species are long-lived and often used in multiple experiments over the course of their lives. Cumulative severity refers to all the positive and negative impacts to health and welfare accumulated across an animal’s lifetime (Pickard, 2013). Thus, the development of biomarkers sensitive to differences in cumulative severity, such as telomere attrition (e.g., Bateson, 2016; Bateson & Poirier, 2019) and changes in hippocampal matter (Bateson & Poirier, 2019; Poirier et al., 2019), is promising.

6.5 Overall conclusions

Measuring animal welfare is a tough but important subject to tackle. Not only are the definitions of good animal welfare constantly under revision, but methods of assessment are also changing. More recently there have been calls for such assessments to take the subject experiences and affective states of animals in account (e.g., Habedank et al., 2018; Kahnau et al., 2020; Lewejohann et al., 2020). As such, the development of welfare and severity assessment tools sensitive to these experiences and states is underway and was the central aim of this thesis. In this thesis, I have proposed the Choice-based Severity Scale and the dot-probe attention bias task as welfare and severity assessment methods that tap into the subjective experiences and affective states of animals. Developing these methods further warrants investigation and refinement particularly as they have the capacity to reform the structure of welfare and severity assessments (i.e., ranking and scaling of welfare parameters) to be objective and reflects the animals’ perspective.

Finding consensus on what welfare parameters are crucial to include in welfare and severity assessments can be difficult, but not impossible. More recently the development of some animal

welfare assessments has been informed by anonymously surveying animal welfare experts about the parameters they judge to be most informative (e.g., cattle: Geist, 2010; Whaytt et al., 2003; laying hens: Whaytt et al., 2003; pigs: Bracke, 2006; Whaytt et al., 2003; horses: Collins et al., 2009; macaques: Truelove et al., 2020; mice: Campos-Luna et al., 2019; Leach et al., 2008; tigers, *Panthera tigris*: Veasey, 2020b; reptiles, *Agamidae*, *Chelidae*, *Pythonidae*, *Testudinidae*: Whittaker et al., 2021; elephants, *Elaphus maximus*: Veasey, 2020a). Those parameters with the highest consensus are generally considered to be the most reliable, practical (i.e., can be assessed in under a day), and valid welfare indices for the species being evaluated. Such considerations are important given that fluctuations in welfare state can occur daily, especially for those settings where we impose conditions upon animals. Assessing expert-identified welfare parameters alongside methods tapping into the psychological experiences of animals is an important next step. As animal welfare scientists, it is ever important that we do not live within the bubble of our discipline but to expand beyond it (Marchant-Forde, 2015). New developments from scientific fields ranging from behavioral ecology to neuroscience will help shape what is to come.

Chapter 7

References

- Abbott, D. H., Keverne, E. B., Bercovitch, F. B., Shively, C. A., Mendoza, S. P., Saltzman, W., Snowdon, C. T., Ziegler, T. E., Banjevic, M., Garland, T., & Sapolsky, R. M. (2003). Are subordinates always stressed? A comparative analysis of rank differences in cortisol levels among primates. *Hormones and Behavior*, *43*, 67–82. [https://doi.org/10.1016/S0018-506X\(02\)00037-5](https://doi.org/10.1016/S0018-506X(02)00037-5)
- Aday, J. S., & Carlson, J. M. (2019). Extended testing with the dot-probe task increases test-retest reliability and validity. *Cognitive Processing*, *20*, 65–72. <https://doi.org/10.1007/s10339-018-0886-1>
- Ahmed, Z., Agha, N., Trunk, A., Berger, M., & Gail, A. (2022). Universal Guide for Skull Extraction and Custom-Fitting of Implants to Continuous and Discontinuous Skulls. *eNeuro*, *9*, ENEURO.0028–22.2022. <https://doi.org/10.1523/ENEURO.0028-22.2022>
- Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. London, UK: Sage Publications.
- Alligood, C. A., Dorey, N. R., Mehrkam, L. R., & Leighty, K. A. (2017). Applying behavior-analytic methodology to the science and practice of environmental enrichment in zoos and aquariums. *Zoo Biology*, *36*, 175–185. <https://doi.org/10.1002/zoo.21368>
- Amir, N., Elias, J., Klumpp, H., & Przeworski, A. (2003). Attentional bias to threat in social phobia: Facilitated processing of threat or difficulty disengaging attention from threat? *Behaviour Research and Therapy*, *41*, 1325–1335. [https://doi.org/10.1016/S0005-7967\(03\)00039-1](https://doi.org/10.1016/S0005-7967(03)00039-1)
- Arce, M., Michopoulos, V., Shepard, K. N., Ha, Q.-C., & Wilson, M. E. (2010). Diet choice, cortisol reactivity, and emotional feeding in socially housed rhesus monkeys. *Physiology & Behavior*, *101*, 446–455. <https://doi.org/10.1016/j.physbeh.2010.07.010>
- Arnold, C. E., & Estep, D. Q. (1994). Laboratory caging preferences in golden hamsters (*Mesocricetus auratus*). *Laboratory Animals*, *28*, 232–238. <https://doi.org/10.1258/002367794780681598>
- Arsenos, G., Hills, J., & Kyriazakis, I. (2000). Conditioned feeding responses of sheep towards flavoured foods associated with casein administration: The role of long delay learning. *Animal Science*, *70*, 157–169. <https://doi.org/10.1017/S1357729800051699>
- Azkona, G., & Sanchez-Pernaute, R. (2022). Mice in translational neuroscience: What R we doing? *Progress in Neurobiology*, *217*, 102330. <https://doi.org/10.1016/j.pneurobio.2022.102330>
- Azzi, R., Fix, D. S., Keller, F. S., & Silva, M. I. R. E. (1964). Exteroceptive control of response

- under delayed reinforcement. *Journal of the Experimental Analysis of Behavior*, 7, 159–162. <https://doi.org/10.1901/jeab.1964.7-159>
- Baker, B. J., & Booth, D. A. (1989). Preference conditioning by concurrent diets with delayed proportional reinforcement. *Physiology & Behavior*, 46, 585–590. [https://doi.org/10.1016/0031-9384\(89\)90336-3](https://doi.org/10.1016/0031-9384(89)90336-3)
- Baker, K. C. (2016). Survey of 2014 behavioral management programs for laboratory primates in the United States. *American Journal of Primatology*, 78, 780–796. <https://doi.org/10.1002/ajp.22543>
- Baker, K. C., Bloomsmith, M. A., Oettinger, B., Neu, K., Griffis, C., & Schoof, V. A. M. (2014). Comparing options for pair housing rhesus macaques using behavioral welfare measures. *American Journal of Primatology*, 76, 30–42. <https://doi.org/10.1002/ajp.22190>
- Baker, K. C., Bloomsmith, M. A., Oettinger, B., Neu, K., Griffis, C., Schoof, V., & Maloney, M. (2012a). Benefits of pair housing are consistent across a diverse population of rhesus macaques. *Applied Animal Behaviour Science*, 137, 148–156. <https://doi.org/10.1016/j.applanim.2011.09.010>
- Baker, K. C., Crockett, C. M., Lee, G. H., Oettinger, B. C., Schoof, V., & Thom, J. P. (2012b). Pair housing for female longtailed and rhesus macaques in the laboratory: Behavior in protected contact versus full contact. *Journal of Applied Animal Welfare Science*, 15, 126–143. <https://doi.org/10.1080/10888705.2012.658330>
- Baker, K. C., Weed, J. L., Crockett, C. M., & Bloomsmith, M. A. (2007). Survey of environmental enhancement programs for laboratory primates. *American Journal of Primatology*, 69, 377–394. <https://doi.org/10.1002/ajp.20347>
- Bala, P. C., Eisenreich, B. R., Yoo, S. B. M., Hayden, B. Y., Park, H. S., & Zimmermann, J. (2020). Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio. *Nature Communications*, 11, 4560. <https://doi.org/10.1038/s41467-020-18441-5>
- Bale, T. L., Abel, T., Akil, H., Carlezon, W. A., Moghaddam, B., Nestler, E. J., Ressler, K. J., & Thompson, S. M. (2019). The critical importance of basic animal research for neuropsychiatric disorders. *Neuropsychopharmacology*, 44, 1349–1353. <https://doi.org/10.1038/s41386-019-0405-9>
- Ballesta, S., Mosher, C. P., Szep, J., Fischl, K. D., & Gothard, K. M. (2016). Social determinants of eyeblinks in adult male macaques. *Scientific Reports*, 6, 38686. <https://doi.org/10.1038/srep38686>
- Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van IJzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin*, 133, 1–24. <https://doi.org/10.1037/0033-2909.133.1.1>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Barry, T. J., Vervliet, B., & Hermans, D. (2015). An integrative review of attention biases

- and their contribution to treatment for anxiety disorders. *Frontiers in Psychology*, 6, 968. <https://doi.org/10.3389/fpsyg.2015.00968>
- Bartlett, J. W., & Frost, C. (2008). Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. *Ultrasound in Obstetrics & Gynecology*, 31, 466–475. <https://doi.org/10.1002/uog.5256>
- Basso, M. A., Frey, S., Guerriero, K. A., Jarraya, B., Kastner, S., Koyano, K. W., Leopold, D. A., Murphy, K., Poirier, C., Pope, W., Silva, A. C., Tansey, G., & Uhrig, L. (2021). Using non-invasive neuroimaging to enhance the care, well-being and experimental outcomes of laboratory non-human primates (monkeys). *NeuroImage*, 228, 117667. <https://doi.org/10.1016/j.neuroimage.2020.117667>
- Bateson, M. (2016). Cumulative stress in research animals: Telomere attrition as a biomarker in a welfare context? *BioEssays*, 38, 201–212. <https://doi.org/10.1002/bies.201500127>
- Bateson, M., & Poirier, C. (2019). Can biomarkers of biological age be used to assess cumulative lifetime experience? *Animal Welfare*, 28, 41–56. <https://doi.org/10.7120/09627286.28.1.041>
- Bayne, K. (2012). Reliance on Behavior as a Metric of Animal Welfare. *ALTEX - Alternatives to Animal Experimentation*, 1, 461–463.
- Behringer, V., Stevens, J. M. G., Hohmann, G., Möstl, E., Selzer, D., & Deschner, T. (2014). Testing the effect of medical positive reinforcement training on salivary cortisol levels in bonobos and orangutans. *PLoS ONE*, 9, e108664. <https://doi.org/10.1371/journal.pone.0108664>
- Bekoff, M. A., & Meaney, C. (1998). *Encyclopedia of animal rights & animal welfare*. Westport, CT, USA: Greenwood Press. Retrieved from <http://site.ebrary.com/lib/roehampton/docDetail.action?docID=5005073>
- Bello, S., Krogsbøll, L. T., Gruber, J., Zhao, Z. J., Fischer, D., & Hróbjartsson, A. (2014). Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias. *Journal of Clinical Epidemiology*, 67, 973–983. <https://doi.org/10.1016/j.jclinepi.2014.04.008>
- Bennett, B. T. (2016). Association of primate veterinarians 2014 nonhuman primate housing survey. *Journal of the American Association for Laboratory Animal Science*, 55, 172–174. Retrieved from <http://www.ingentaconnect.com/content/aalas/jaalas/2016/00000055/00000002/art00008>
- Berger, M., Calapai, A., Stephan, V., Niessing, M., Burchardt, L., Gail, A., & Treue, S. (2018). Standardized automated training of rhesus monkeys for neuroscience research in their housing environment. *Journal of Neurophysiology*, 119, 796–807. <https://doi.org/10.1152/jn.00614.2017>
- Bernstein, I. L. (1999). Taste aversion learning: A contemporary perspective. *Nutrition*, 15, 229–234. [https://doi.org/10.1016/S0899-9007\(98\)00192-0](https://doi.org/10.1016/S0899-9007(98)00192-0)
- Bernstein, I. S., Weed, J. L., Judge, P. G., & Ruehlmann, T. E. (1989). Seasonal weight changes in male rhesus monkeys (*Macaca mulatta*). *American Journal of Primatology*, 18, 251–257. <https://doi.org/10.1002/ajp.1350180309>

- Bethell, E. J. (2015). A “how-to” guide for designing judgment bias studies to assess captive animal welfare. *Journal of Applied Animal Welfare Science*, *18*, S18–S42. <https://doi.org/10.1080/10888705.2015.1075833>
- Bethell, E. J., Cassidy, L. C., Brockhausen, R. R., & Pfefferle, D. (2019). Toward a standardized test of fearful temperament in primates: A sensitive alternative to the human intruder task for laboratory-housed rhesus macaques (*Macaca mulatta*). *Frontiers in Psychology*, *10*, 1051. <https://doi.org/10.3389/fpsyg.2019.01051>
- Bethell, E. J., Holmes, A., MacLarnon, A., & Semple, S. (2012a). Evidence that emotion mediates social attention in rhesus macaques. *PLoS ONE*, *7*, e44387. <https://doi.org/10.1371/journal.pone.0044387>
- Bethell, E., Holmes, A., Maclarnon, A., & Semple, S. (2012b). Cognitive bias in a non-human primate: Husbandry procedures influence cognitive indicators of psychological well-being in captive rhesus macaques. *Animal Welfare*, *21*, 185–195. <https://doi.org/10.7120/09627286.21.2.185>
- Bethell, E., Holmes, A., MacLarnon, A., & Semple, S. (2016). Emotion evaluation and response slowing in a non-human primate: New directions for cognitive bias measures of animal emotion? *Behavioral Sciences*, *6*, 2. <https://doi.org/10.3390/bs6010002>
- Bicca-Marques, J. C., & Garber, P. A. (2004). Use of spatial, visual, and olfactory information during foraging in wild nocturnal and diurnal anthropoids: A field experiment comparing *Aotus*, *Callicebus*, and *Saguinus*. *American Journal of Primatology*, *62*, 171–187. <https://doi.org/10.1002/ajp.20014>
- Blanchard, T. C., & Hayden, B. Y. (2015). Monkeys are more patient in a foraging task than in a standard intertemporal choice task. *PLoS ONE*, *10*, e0117057. <https://doi.org/10.1371/journal.pone.0117057>
- Blanchard, T. C., Pearson, J. M., & Hayden, B. Y. (2013). Postreward delays and systematic biases in measures of animal temporal discounting. *Proceedings of the National Academy of Sciences*, *110*, 15491–15496. <https://doi.org/10.1073/pnas.1310446110>
- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, *327*, 307–310. <https://doi.org/10.1016/j.ijnurstu.2009.10.001>
- Bliss-Moreau, E., Theil, J. H., & Moadab, G. (2013). Efficient cooperative restraint training with rhesus macaques. *Journal of Applied Animal Welfare Science*, *16*, 98–117. <https://doi.org/10.1080/10888705.2013.768897>
- Boissy, A., Manteuffel, G., Jensen, M. B., Moe, R. O., Spruijt, B., Keeling, L. J., Winckler, C., Forkman, B., Dimitrov, I., Langbein, J., Bakken, M., Veissier, I., & Aubert, A. (2007). Assessment of positive emotions in animals to improve their welfare. *Physiology & Behavior*, *92*, 375–397. <https://doi.org/10.1016/j.physbeh.2007.02.003>
- Botreau, R., Veissier, I., & Perny, P. (2009). Overall assessment of animal welfare: Strategy adopted in Welfare Quality®. *Animal Welfare*, *18*, 363–370.
- Botvinick, M. M., Huffstetler, S., & McGuire, J. T. (2009). Effort discounting in human nucleus accumbens. *Cognitive, Affective, & Behavioral Neuroscience*, *9*, 16–27. <https://doi.org/10.1007/s12027-009-9100-0>

- 3758/CABN.9.1.16
- Bracke, M. (2006). Expert opinion regarding environmental enrichment materials for pigs. *Animal Welfare, 15*, 67–70.
- Bradley, B. P., Mogg, K., Falla, S. J., & Hamilton, L. R. (1998). Attentional bias for threatening facial expressions in anxiety: Manipulation of stimulus duration. *Cognition and Emotion, 12*, 737–753. <https://doi.org/10.1080/026999398379411>
- Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*, 602–607. <https://doi.org/10.1111/j.1469-8986.2008.00654.x>
- Brambell, R. (1965). *Report of the technical committee to enquire into the welfare of animals kept under intensive livestock husbandry systems*. London, UK: Her Majesty's Stationery Office.
- Broom, D. (2007). Quality of life means welfare: How is it related to other concepts and assessed? *Animal Welfare, 16*, 45–53.
- Broom, D. M. (1988). The scientific assessment of animal welfare. *Applied Animal Behaviour Science, 20*, 5–19. [https://doi.org/10.1016/0168-1591\(88\)90122-0](https://doi.org/10.1016/0168-1591(88)90122-0)
- Broom, D. M. (2014). *Sentience and animal welfare*. Oxfordshire, UK: CAB International.
- Browning, H. (2020). Assessing Measures of Animal Welfare. *PhilSci-Archive*. Retrieved from <http://philsci-archive.pitt.edu/id/eprint/17144>
- Browning, H., & Birch, J. (2022). Animal sentience. *Philosophy Compass, 17*, e12822. <https://doi.org/10.1111/phc3.12822>
- Browning, H., & Veit, W. (2022). The sentience shift in animal research. *The New Bioethics, 1–16*. <https://doi.org/10.1080/20502877.2022.2077681>
- Bruce, V., & Young, A. (2012). *Face perception*. Psychology Press.
- Buchanan-Smith, H. M., & Badihi, I. (2012). The psychology of control: Effects of control over supplementary light on welfare of marmosets. *Applied Animal Behaviour Science, 137*, 166–174. <https://doi.org/10.1016/j.applanim.2011.07.002>
- Buchanan-Smith, H., Rennie, A., Vitale, A., Pollo, S., Prescott, M., & Morton, D. (2005). Harmonising the definition of refinement. *Animal Welfare, 14*, 379–384.
- Bugnon, P., Heimann, M., & Thallmair, M. (2016). What the literature tells us about score sheet design. *Laboratory Animals, 50*, 414–417. <https://doi.org/10.1177/0023677216671552>
- Burdick, N. C., Randel, R. D., Carroll, J. A., & Welsh, T. H. (2011). Interactions between temperament, stress, and immune function in cattle. *International Journal of Zoology, 2011*, e373197. <https://doi.org/10.1155/2011/373197>
- Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*, 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Burman, O. H. P., Parker, R. M. A., Paul, E. S., & Mendl, M. T. (2009). Anxiety-induced cognitive bias in non-human animals. *Physiology & Behavior, 98*, 345–350. <https://doi.org/10.1016/j.physbeh.2009.06.012>
- Calapai, A., Berger, M., Niessing, M., Heisig, K., Brockhausen, R., Treue, S., & Gail, A. (2017). A cage-based training, cognitive testing and enrichment system optimized for rhesus

- macaques in neuroscience research. *Behavior Research Methods*, *49*, 35–45. <https://doi.org/10.3758/s13428-016-0707-3>
- Calapai, A., Cabrera-Moreno, J., Moser, T., & Jeschke, M. (2022). Flexible auditory training, psychophysics, and enrichment of common marmosets with an automated, touchscreen-based system. *Nature Communications*, *13*, 1648. <https://doi.org/10.1038/s41467-022-29185-9>
- Calder, A. J., Young, A. W., Perrett, D. I., Ectoff, N. L., & Rowland, D. (1996). Categorical Perception of Morphed Facial Expressions. *Visual Cognition*, *3*, 81–118. <https://doi.org/10.1080/713756735>
- Campos-Luna, I., Miller, A., Beard, A., & Leach, M. (2019). Validation of mouse welfare indicators: A Delphi consultation survey. *Scientific Reports*, *9*, 10249. <https://doi.org/10.1038/s41598-019-45810-y>
- Capaldi, E. D., Campbell, D. H., Sheffer, J. D., & Bradford, J. P. (1987). Conditioned flavor preferences based on delayed caloric consequences. *Journal of Experimental Psychology: Animal Behavior Processes*, *13*, 150–155. <https://doi.org/10.1037/0097-7403.13.2.150>
- Capitano, J. P., & Emborg, M. E. (2008). Contributions of non-human primates to neuroscience research. *The Lancet*, *371*, 1126–1135. [https://doi.org/10.1016/S0140-6736\(08\)60489-4](https://doi.org/10.1016/S0140-6736(08)60489-4)
- Capitano, J. P., Kyes, R. C., & Fairbanks, L. A. (2006). Considerations in the selection and conditioning of Old World monkeys for laboratory research: Animals from domestic sources. *ILAR Journal*, *47*, 294–306. <https://doi.org/10.1093/ilar.47.4.294>
- Carbone, L. (2021). Estimating mouse and rat use in American laboratories by extrapolation from Animal Welfare Act-regulated species. *Scientific Reports*, *11*, 493. <https://doi.org/10.1038/s41598-020-79961-0>
- Cardinal, R. N., Parkinson, J. A., Hall, J., & Everitt, B. J. (2002). Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience & Biobehavioral Reviews*, *26*, 321–352. [https://doi.org/10.1016/S0149-7634\(02\)00007-6](https://doi.org/10.1016/S0149-7634(02)00007-6)
- Carp, S. B., Rothwell, E. S., Bourdon, A., Freeman, S. M., Ferrer, E., & Bales, K. L. (2016). Development of a partner preference test that differentiates between established pair bonds and other relationships in socially monogamous titi monkeys (*Callicebus cupreus*). *American Journal of Primatology*, *78*, 326–339. <https://doi.org/10.1002/ajp.22450>
- Carr, A. R., Scully, A., Webb, M., & Felmingham, K. L. (2016). Gender differences in salivary alpha-amylase and attentional bias towards negative facial expressions following acute stress induction. *Cognition and Emotion*, *30*, 315–324. <https://doi.org/10.1080/02699931.2014.999748>
- Cassidy, L. C., Bethell, E. J., Brockhausen, R. R., Boretius, S., Treue, S., & Pfefferle, D. (2021). The dot-probe attention bias task as a method to assess psychological wellbeing after anesthesia: A study with adult female long-tailed macaques (*Macaca fascicularis*). *European Surgical Research*. <https://doi.org/10.1159/000521440>
- Cassidy, L. C., Hannibal, D. L., Semple, S., & McCowan, B. (2020). Improved behavioral indices of welfare in continuous compared to intermittent pair housing in adult female rhesus macaques (*Macaca mulatta*). *American Journal of Primatology*, *82*, e23189. <https://doi.org/10.1002/ajp.22450>

- [//doi.org/10.1002/ajp.23189](https://doi.org/10.1002/ajp.23189)
- Charbonneau, J. A., Amaral, D. G., & Bliss-Moreau, E. (2021). Social housing status impacts rhesus monkeys' affective responding in classic threat processing tasks. *bioRxiv*. <https://doi.org/10.1101/2021.05.16.444352>
- Cisler, J. M., Bacon, A. K., & Williams, N. L. (2009). Phenomenological characteristics of attentional biases towards threat: A critical review. *Cognitive Therapy and Research, 33*, 221–234. <https://doi.org/10.1007/s10608-007-9161-y>
- Cisler, J. M., & Koster, E. H. W. (2010). Mechanisms of attentional biases towards threat in anxiety disorders: An integrative review. *Clinical Psychology Review, 30*, 203–216. <https://doi.org/10.1016/j.cpr.2009.11.003>
- Clarence, W. M., Scott, J. P., Dorris, M. C., & Paré, M. (2006). Use of enclosures with functional vertical space by captive rhesus monkeys (*Macaca mulatta*) involved in biomedical research. *Journal of the American Association for Laboratory Animal Science, 45*, 31–34.
- Clark, F. (2017). Cognitive enrichment and welfare: Current approaches and future directions. *Animal Behavior and Cognition, 4*, 52–71. <https://doi.org/10.12966/abc.05.02.2017>
- Clark, F. E. (2011). Great ape cognition and captive care: Can cognitive challenges enhance well-being? *Applied Animal Behaviour Science, 135*, 1–12. <https://doi.org/10.1016/j.applanim.2011.10.010>
- Clarke, T., Pluske, J. R., & Fleming, P. A. (2016). Are observer ratings influenced by prescription? A comparison of free choice profiling and fixed list methods of qualitative behavioural assessment. *Applied Animal Behaviour Science, 177*, 77–83. <https://doi.org/10.1016/j.applanim.2016.01.022>
- Clegg, I. L. K. (2018). Cognitive bias in zoo animals: An optimistic outlook for welfare assessment. *Animals, 8*, 104. <https://doi.org/10.3390/ani8070104>
- Clegg, I. L. K., Rödel, H. G., Mercera, B., van der Heul, S., Schrijvers, T., de Laender, P., Gojceta, R., Zimmitti, M., Verhoeven, E., Burger, J., Bunschoek, P. E., & Delfour, F. (2019). Dolphins' willingness to participate (WtP) in positive reinforcement training as a potential welfare indicator, where WtP predicts early changes in health status. *Frontiers in Psychology, 10*, 2112. <https://doi.org/10.3389/fpsyg.2019.02112>
- Cohen, S., & Beths, T. (2020). Grimace scores: Tools to support the identification of pain in mammals used in research. *Animals, 10*, 1726. <https://doi.org/10.3390/ani10101726>
- Coleman, K. (2012). Individual differences in temperament and behavioral management practices for nonhuman primates. *Applied Animal Behaviour Science, 137*, 106–113. <https://doi.org/10.1016/j.applanim.2011.08.002>
- Collins, J., Hanlon, A., More, S. J., Wall, P. G., & Duggan, V. (2009). Policy Delphi with vignette methodology as a tool to evaluate the perception of equine welfare. *The Veterinary Journal, 181*, 63–69. <https://doi.org/10.1016/j.tvjl.2009.03.012>
- Cook, C., Mellor, D., Harris, P., Ingram, J., & Matthews, L. (2000). Hands-on and hands-off measurement of stress. In G. P. Moberg & J. A. Mench (Eds.), *The biology of animal stress: Basic principles and implications for animal welfare* (pp. 123–46). New York: CABI Publishing.

- Cooper, J. (2004). Consumer demand under commercial husbandry conditions: Practical advice on measuring behavioural priorities in captive animals. *Animal Welfare*, *13*, 47–56.
- Cooper, R. M., & Langton, S. R. H. (2006). Attentional bias to angry faces using the dot-probe task? It depends when you look for it. *Behaviour Research and Therapy*, *44*, 1321–1329. <https://doi.org/10.1016/j.brat.2005.10.004>
- Costa, C. S., Oliveira, A. W. C., Easton, A., & Barros, M. (2022). A single brief stressful event time-dependently affects object recognition memory and promotes familiarity preference in marmoset monkeys. *Behavioural Processes*, *199*, 104645. <https://doi.org/10.1016/j.beproc.2022.104645>
- Cox, L. A., Olivier, M., Spradling-Reeves, K., Karere, G. M., Comuzzie, A. G., & VandeBerg, J. L. (2017). Nonhuman primates and translational research—Cardiovascular disease. *ILAR Journal*, *58*, 235–250. <https://doi.org/10.1093/ilar/ilx025>
- Cronin, K. A., Bethell, E. J., Jacobson, S. L., Egelkamp, C., & Hopper, L. M. (2018). Evaluating mood changes in response to anthropogenic noise with a response-slowing task in three species of zoo-housed primates. *Animal Behavior and Cognition*, *5*, 209–221. <https://doi.org/10.26451/abc.05.02.03.2018>
- Crump, A., Arnott, G., & Bethell, E. (2018). Affect-driven attention biases as animal welfare indicators: Review and methods. *Animals*, *8*, 136. <https://doi.org/10.3390/ani8080136>
- Crump, A., Bethell, E. J., Earley, R., Lee, V. E., Mendl, M., Oldham, L., Turner, S. P., & Arnott, G. (2020). Emotion in animal contests. *Proceedings of the Royal Society B: Biological Sciences*, *287*, 20201715. <https://doi.org/10.1098/rspb.2020.1715>
- D’Amato, M. R., & Puopolo, M. (1981). Long-delay spatial discrimination learning in monkeys (*Cebus apella*). *Bulletin of the Psychonomic Society*, *18*, 85–88. <https://doi.org/10.3758/BF03333567>
- D’Amato, M. R., Salmon, D. P., & Puopolo, M. (1981). Long-delay visual discrimination learning in monkeys (*Cebus apella*). *Bulletin of the Psychonomic Society*, *18*, 89–91. <https://doi.org/10.3758/BF03333568>
- Dawkins, M. (1977). Do hens suffer in battery cages? Environmental preferences and welfare. *Animal Behaviour*, *25*, 1034–1046. [https://doi.org/10.1016/0003-3472\(77\)90054-9](https://doi.org/10.1016/0003-3472(77)90054-9)
- Dawkins, M. (1980). *Animal suffering: The science of animal welfare*. New York, United States: Chapman and Hall.
- Dawkins, M. S. (1998). Evolution and animal welfare. *Quarterly Review of Biology*, 305–328. <https://doi.org/10.1086/420307>
- Dawkins, M. S. (2006). A user’s guide to animal welfare science. *Trends in Ecology & Evolution*, *21*, 77–82. <https://doi.org/10.1016/j.tree.2005.10.017>
- Dawkins, M. S., & Beardsley, T. (1986). Reinforcing properties of access to litter in hens. *Applied Animal Behaviour Science*, *15*, 351–364. [https://doi.org/10.1016/0168-1591\(86\)90127-9](https://doi.org/10.1016/0168-1591(86)90127-9)
- de Boyer des Roches, A., Lussert, A., Faure, Marion., Herry, Vincent., Rainard, Pascal., Durand, Denys., Wemelsfelder, F., & Foucras, G. (2018). Dairy cows under experimentally-induced *Escherichia coli* mastitis show negative emotional states assessed through Qualitative

- Behaviour Assessment. *Applied Animal Behaviour Science*, 206, 1–11. <https://doi.org/10.1016/j.applanim.2018.06.004>
- de Vere, A. J., & Kuczaj, S. A. (2016). Where are we in the study of animal emotions? *Wiley Interdisciplinary Reviews: Cognitive Science*, 7, 354–362. <https://doi.org/10.1002/wcs.1399>
- de Vries, R. B. M., Hooijmans, C. R., Tillema, A., Leenaars, M., & Ritskes-Hoitinga, M. (2011). A search filter for increasing the retrieval of animal studies in Embase. *Laboratory Animals*, 45, 268–270. <https://doi.org/10.1258/la.2011.011056>
- de Vries, R. B. M., Hooijmans, C. R., Tillema, A., Leenaars, M., & Ritskes-Hoitinga, M. (2014). Updated version of the Embase search filter for animal studies. *Laboratory Animals*, 48, 88–88. <https://doi.org/10.1177/0023677213494374>
- Descovich, K. (2017). Facial expression: An under-utilised tool for the assessment of welfare in mammals. *ALTEX - Alternatives to Animal Experimentation*, 34, 409–429. <https://doi.org/10.14573/altex.1607161>
- Descovich, K. A., Richmond, S. E., Leach, M. C., Buchanan-Smith, H. M., Flecknell, P., Farningham, D. A. H., Witham, C., Gates, M. C., & Vick, S.-J. (2019). Opportunities for refinement in neuroscience: Indicators of wellness and post-operative pain in laboratory macaques. *ALTEX - Alternatives to Animal Experimentation*, 36, 535–554. <https://doi.org/10.14573/altex.1811061>
- Dominy, N. J., Lucas, P. W., Osorio, D., & Yamashita, N. (2001). The sensory ecology of primate food perception. *Evolutionary Anthropology: Issues, News, and Reviews*, 10, 171–186. <https://doi.org/10.1002/evan.1031>
- Doyle, R. E., Vidal, S., Hinch, G. N., Fisher, A. D., Boissy, A., & Lee, C. (2010). The effect of repeated testing on judgement biases in sheep. *Behavioural Processes*, 83, 349–352. <https://doi.org/10.1016/j.beproc.2010.01.019>
- Duncan, I. J. (1978). The interpretation of preference tests in animal behaviour. *Applied Animal Ethology*, 4, 197–200. [https://doi.org/10.1016/0304-3762\(78\)90086-X](https://doi.org/10.1016/0304-3762(78)90086-X)
- Duncan, I. J. H. (1992). Measuring Preferences and the Strength of Preferences. *Poultry Science*, 71, 658–663. <https://doi.org/10.3382/ps.0710658>
- Egelkamp, C. L., & Ross, S. R. (2019). A review of zoo-based cognitive research using touchscreen interfaces. *Zoo Biology*, 38, 220–235. <https://doi.org/10.1002/zoo.21458>
- Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, 44, 227–240. [https://doi.org/10.1016/0010-0277\(92\)90002-Y](https://doi.org/10.1016/0010-0277(92)90002-Y)
- European Commission. (2020). 2019 report on the statistics on the use of animals for scientific purposes in the Member States of the European Union in 2015-2017. Retrieved from [https://ec.europa.eu/transparency/documents-register/detail?ref=COM\(2020\)16&lang=en](https://ec.europa.eu/transparency/documents-register/detail?ref=COM(2020)16&lang=en)
- European Parliament. (2022). EU directive 2010/63/EU of the European parliament and of the council of 22 September 2010 on the protection of animals used for scientific purposes. Retrieved from <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:276:0033:0079:en:PDF>
- Evans, T. A., & Beran, M. J. (2007). Delay of gratification and delay maintenance by rhesus

- macaques (*Macaca mulatta*). *The Journal of General Psychology*, *134*, 199–216. <https://doi.org/10.3200/GENP.134.2.199-216>
- Everds, N. E., Snyder, P. W., Bailey, K. L., Bolon, B., Creasy, D. M., Foley, G. L., Rosol, T. J., & Sellers, T. (2013). Interpreting stress responses during routine toxicity studies: A review of the biology, impact, and assessment. *Toxicologic Pathology*, *41*, 560–614. <https://doi.org/10.1177/0192623312466452>
- Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *NeuroReport*, *9*, 303–308.
- Fagot, J., Martin-Malivel, J., & Dépy, D. (2000). What is the evidence for an equivalence between objects and pictures in birds and nonhuman primates? In J. Fagot (Ed.), *Picture perception in animals* (pp. 295–320). Philadelphia, PA: Psychology Press.
- Fagot, J., Thompson, R. K. R., & Parron, C. (2010). How to read a picture: Lessons from nonhuman primates. *Proceedings of the National Academy of Sciences*, *107*, 519–520. <https://doi.org/10.1073/pnas.0913577107>
- Fernandez, E. J., Dorey, N., & Rosales-Ruiz, J. (2004). A two-choice preference assessment with five cotton-top tamarins (*Saguinus oedipus*). *Journal of Applied Animal Welfare Science*, *7*, 163–169. https://doi.org/10.1207/s15327604jaws0703_2
- Fernandez, E. J., & Timberlake, W. (2019). Selecting and testing environmental enrichment in lemurs. *Frontiers in Psychology*, *10*, 2119. <https://doi.org/10.3389/fpsyg.2019.02119>
- Ferster, C. B., & Hammer, C. (1965). Variables determining the effects of delay in reinforcement. *Journal of the Experimental Analysis of Behavior*, *8*, 243–254. <https://doi.org/10.1901/jeab.1965.8-243>
- Fox, E., Russo, R., Bowles, R., & Dutton, K. (2001). Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of Experimental Psychology: General*, *130*, 681–700. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1924776/>
- Fraser, D. (1999). Animal ethics and animal welfare science: Bridging the two cultures. *Applied Animal Behaviour Science*, *65*, 171–189. [https://doi.org/10.1016/S0168-1591\(99\)00090-8](https://doi.org/10.1016/S0168-1591(99)00090-8)
- Fraser, D. (2009). Assessing animal welfare: Different philosophies, different scientific approaches. *Zoo Biology*, *28*, 507–518.
- Fraser, D., & Matthews, L. R. (1997). Preference and motivation testing. In M. C. Appleby & B. O. Hughes (Eds.), *Animal welfare* (pp. 159–173). Wallingford, UK: CAB International.
- Fraser, D., & Nicol, C. J. (2011). Preference and motivation research. In M. C. Appleby, J. A. Mench, I. A. S. Olson, & B. O. Hughes (Eds.), *Animal welfare* (pp. 183–199). Wallingford, UK: CAB International.
- Freymann, J., Tsai, P.-P., Stelzer, H. D., Mischke, R., & Hackbarth, H. (2017). Impact of bedding volume on physiological and behavioural parameters in laboratory mice. *Laboratory Animals*, *51*, 601–612. <https://doi.org/10.1177/0023677217694400>
- Freymann, J., Tsai, P.-P., Stelzer, H., & Hackbarth, H. (2015). The amount of cage bedding preferred by female BALB/c and C57BL/6 mice. *Lab Animal*, *44*, 17–22. <https://doi.org/10.1038/lab.659>
- Froesel, M., Goudard, Q., Hauser, M., Gacoin, M., & Ben Hamed, S. (2020). Automated

- video-based heart rate tracking for the anesthetized and behaving monkey. *Scientific Reports*, *10*, 17940. <https://doi.org/10.1038/s41598-020-74954-5>
- Gabry, J., & Mahr, T. (2022). Package 'bayesplot': Plotting for Bayesian models (Version R package version 1.9.0). R package version 1.9.0. Retrieved from <https://mc-stan.org/bayesplot/>
- Garber, P. A. (2000). Evidence for the use of spatial, temporal, and social information by some primate foragers. In S. Boinski & P. A. Garber (Eds.), *On the move: How and why animals travel in groups* (pp. 261–298). Chicago: Chicago University Press.
- Gartner, M. C., & Weiss, A. (2013). Scottish wildcat (*Felis silvestris grampia*) personality and subjective well-being: Implications for captive management. *Applied Animal Behaviour Science*, *147*, 261–267. <https://doi.org/10.1016/j.applanim.2012.11.002>
- Gatti, E., Calzolari, E., Maggioni, E., & Obrist, M. (2018). Emotional ratings and skin conductance response to visual, auditory and haptic stimuli. *Scientific Data*, *5*, 180120. <https://doi.org/10.1038/sdata.2018.120>
- Geist, M. R. (2010). Using the Delphi method to engage stakeholders: A comparison of two studies. *Evaluation and Program Planning*, *33*, 147–154. <https://doi.org/10.1016/j.evalprogplan.2009.06.006>
- Genty, E., Karpel, H., & Silberberg, A. (2012). Time preferences in long-tailed macaques (*Macaca fascicularis*) and humans (*Homo sapiens*). *Animal Cognition*, *15*, 1161–1172. <https://doi.org/10.1007/s10071-012-0540-8>
- Gilbert, M. H., & Baker, K. C. (2011). Social buffering in adult male rhesus macaques (*Macaca mulatta*): Effects of stressful events in single vs. Pair housing. *Journal of Medical Primatology*, *40*, 71–78. <https://doi.org/10.1111/j.1600-0684.2010.00447.x>
- Goodman, S., & Check, E. (2002). Animal experiments: The great primate debate. *Nature*, *417*, 684–687.
- Gottlieb, D. H., Capitanio, J. P., & McCowan, B. (2013a). Risk factors for stereotypic behavior and self-biting in rhesus macaques (*Macaca Mulatta*): Animal's history, current environment, and personality. *American Journal of Primatology*, *75*, 995–1008. <https://doi.org/10.1002/ajp.22161>
- Gottlieb, D. H., Coleman, K., & McCowan, B. (2013b). The effects of predictability in daily husbandry routines on captive rhesus macaques (*Macaca mulatta*). *Applied Animal Behaviour Science*, *143*, 117–127. <https://doi.org/10.1016/j.applanim.2012.10.010>
- Gottlieb, D. H., Maier, A., & Coleman, K. (2015). Evaluation of environmental and intrinsic factors that contribute to stereotypic behavior in captive rhesus macaques (*Macaca Mulatta*). *Applied Animal Behaviour Science*, *171*, 184–191. <https://doi.org/10.1016/j.applanim.2015.08.005>
- Goymann, W. (2012). On the Use of non-invasive hormone research in uncontrolled, natural environments: The problem with sex, diet, metabolic rate and the individual. *Methods in Ecology and Evolution*, *3*, 757–765. <https://doi.org/10.1111/j.2041-210X.2012.00203.x>
- Grames, E. M., Stillman, A. N., Tingley, M. W., & Elphick, C. S. (2019). An automated approach to identifying search terms for systematic reviews using keyword co-occurrence

- networks. *Methods in Ecology and Evolution*, *10*, 1645–1654. <https://doi.org/10.1111/2041-210X.13268>
- Gray, H., Bertrand, H., Mindus, C., Flecknell, P., Rowe, C., & Thiele, A. (2016). Physiological, behavioral, and scientific impact of different fluid control protocols in the rhesus macaque (*Macaca mulatta*). *eNeuro*, *3*, ENEURO.0195–16.2016. <https://doi.org/10.1523/ENEURO.0195-16.2016>
- Gray, H., Pearce, B., Thiele, A., & Rowe, C. (2017). The use of preferred social stimuli as rewards for rhesus macaques in behavioural neuroscience. *PLoS ONE*, *12*, e0178048. <https://doi.org/10.1371/journal.pone.0178048>
- Gray, H., Thiele, A., & Rowe, C. (2019). Using preferred fluids and different reward schedules to motivate rhesus macaques (*Macaca mulatta*) in cognitive tasks. *Laboratory Animals*, *53*, 372–382. <https://doi.org/10.1177/0023677218801390>
- Green, T., & Mellor, D. (2011). Extending ideas about animal welfare assessment to include “quality of life” and related concepts. *New Zealand Veterinary Journal*, *59*, 263–271. <https://doi.org/10.1080/00480169.2011.610283>
- Gygax, L. (2014). The A to Z of statistics for testing cognitive judgement bias. *Animal Behaviour*, *95*, 59–69. <https://doi.org/10.1016/j.anbehav.2014.06.013>
- Habedank, A., Kahnau, P., Diederich, K., & Lewejohann, L. (2018). Severity assessment from an animal’s point of view. *Berliner Und Münchener Tierärztliche Wochenschrift*, *131*, 304–320. Retrieved from <https://www.cabdirect.org/cabdirect/abstract/20183235031>
- Habedank, A., Kahnau, P., & Lewejohann, L. (2021). Alternate without alternative: Neither preference nor learning explains behaviour of C57BL/6J mice in the T-maze. *Behaviour*, *158*, 625–662. <https://doi.org/10.1163/1568539X-bja10085>
- Haddaway, N. R., Woodcock, P., Macura, B., & Collins, A. (2015). Making literature reviews more reliable through application of lessons from systematic reviews. *Conservation Biology*, *29*, 1596–1605. <https://doi.org/10.1111/cobi.12541>
- Hannibal, D. L., Bliss-Moreau, E., Vandeleest, J., McCowan, B., & Capitanio, J. (2017). Laboratory rhesus macaque social housing and social changes: Implications for research. *American Journal of Primatology*, *79*, e22528. <https://doi.org/10.1002/ajp.22528>
- Hannibal, D. L., Cassidy, L. C., Vandeleest, J., Semple, S., Barnard, A., Chun, K., Winkler, S., & McCowan, B. (2018). Intermittent pair-housing, pair relationship qualities, and HPA activity in adult female rhesus macaques. *American Journal of Primatology*, *80*, e22762. <https://doi.org/10.1002/ajp.22762>
- Hansell, M., Åsberg, A., & Laska, M. (2020). Food preferences and nutrient composition in zoo-housed ring-tailed lemurs, *Lemur catta*. *Physiology & Behavior*, *226*, 113125. <https://doi.org/10.1016/j.physbeh.2020.113125>
- Harding, E. J., Paul, E. S., & Mendl, M. (2004). Cognitive bias and affective state. *Nature*, *427*, 312–312. <https://doi.org/10.1038/427312a>
- Harrison, R. (1964). *Animal machines*. London, UK: Vincent Stuart Ltd.
- Hau, M., & Goymann, W. (2015). Endocrine mechanisms, behavioral phenotypes and plasticity: Known relationships and open questions. *Frontiers in Zoology*, *12*, S7.

- <https://doi.org/10.1186/1742-9994-12-S1-S7>
- Hausner, E., Waffenschmidt, S., Kaiser, T., & Simon, M. (2012). Routine development of objectively derived search strategies. *Systematic Reviews*, *1*, 19. <https://doi.org/10.1186/2046-4053-1-19>
- Hawkins, P., Morton, D., Burman, O., Dennison, N., Honess, P., Jennings, M., Lane, S., Middleton, V., Roughan, J., & Wells, S. (2011). A guide to defining and implementing protocols for the welfare assessment of laboratory animals: Eleventh report of the BVAWF/FRAME/RSPCA/UFAW Joint Working Group on Refinement. *Laboratory Animals*, *45*, 1–13.
- Hayden, B. Y. (2016). Time discounting and time preference in animals: A critical review. *Psychonomic Bulletin & Review*, *23*, 39–53. <https://doi.org/10.3758/s13423-015-0879-3>
- Held, S. D. E., & Špinka, M. (2011). Animal play and animal welfare. *Animal Behaviour*, *81*, 891–899. <https://doi.org/10.1016/j.anbehav.2011.01.007>
- Henderson, R. R., Bradley, M. M., & Lang, P. J. (2018). Emotional imagery and pupil diameter. *Psychophysiology*, *55*, e13050. <https://doi.org/10.1111/psyp.13050>
- Hennessy, M. B., Kaiser, S., & Sachser, N. (2009). Social buffering of the stress response: Diversity, mechanisms, and functions. *Frontiers in Neuroendocrinology*, *30*, 470–482. <https://doi.org/10.1016/j.yfrne.2009.06.001>
- Hernández, M. C., González-Campos, S., & Barja, I. (2021). Colour preferences in relation to diet in chimpanzees (*Pan troglodytes*), gorillas (*Gorilla gorilla*) and mandrills (*Mandrillus sphinx*). *Folia Primatologica*, *92*, 306–314. <https://doi.org/10.1159/000520487>
- Hobbiesiefken, U., Urmersbach, B., Jaap, A., Diederich, K., & Lewejohann, L. (2021). Rating enrichment items by group-housed laboratory mice in multiple binary choice tests using an RFID-based tracking system. *PLoS ONE*, *16*, e0261876. <https://doi.org/10.1101/2021.10.20.465117>
- Holmes, A., Winston, J. S., & Eimer, M. (2005). The role of spatial frequency information for ERP components sensitive to faces and emotional facial expression. *Cognitive Brain Research*, *25*, 508–520. <https://doi.org/10.1016/j.cogbrainres.2005.08.003>
- Homberg, J. R., Adan, R. A. H., Alenina, N., Asiminas, A., Bader, M., Beckers, T., ... Genzel, L. (2021). The continued need for animals to advance brain research. *Neuron*, *109*, 2374–2379. <https://doi.org/10.1016/j.neuron.2021.07.015>
- Home Office. (2019). Statistics of Scientific Procedures on Living Animals, Great Britain 2019. Her Majesties Stationary Office, London, UK. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/901224/annual-statistics-scientific-procedures-living-animals-2019.pdf
- Honess, P., & Wolfensohn, S. (2010). The extended welfare assessment grid: A matrix for the assessment of welfare and cumulative suffering in experimental animals. *ATLA-Alternatives to Laboratory Animals*, *38*, 205.
- Hooijmans, C. R., Tillema, A., Leenaars, M., & Ritskes-Hoitinga, M. (2010). Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. *Laboratory Animals*, *44*, 170–175. <https://doi.org/10.1258/la.2010.009117>

- Hopper, L. M. (2017). Cognitive research in zoos. *Current Opinion in Behavioral Sciences*, *16*, 100–110. <https://doi.org/10.1016/j.cobeha.2017.04.006>
- Hopper, L. M. (2022). Primatology in zoos: Studying behavior, cognition, and welfare. *American Journal of Primatology*, e23385. <https://doi.org/10.1002/ajp.23385>
- Hopper, L. M., Gulli, R. A., Howard, L. H., Kano, F., Krupenye, C., Ryan, A. M., & Paukner, A. (2020). The application of noninvasive, restraint-free eye-tracking methods for use with nonhuman primates. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-020-01465-6>
- Hosey, G. R., Jacques, M., & Burton, M. (1999). Allowing captive marmosets to choose the size and position of their nest box. *Animal Welfare*, *8*, 281–285. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1999-03665-007&site=ehost-live>
- Howarth, E. R., Kemp, C., Thatcher, H., Szott, I. D., Farningham, D., Whitham, C. L., Holmes, A., Semple, S., & Bethell, E. J. (2021). Developing and validating attention bias tools for assessing trait and state affect in animals: A worked example with *Macaca mulatta*. *Applied Animal Behaviour Science*, *234*, 105198.
- Howell, B. R., Ahn, M., Shi, Y., Godfrey, J. R., Hu, X., Zhu, H., Styner, M., & Sanchez, M. M. (2019). Disentangling the effects of early caregiving experience and heritable factors on brain white matter development in rhesus monkeys. *NeuroImage*, *197*, 625–642. <https://doi.org/10.1016/j.neuroimage.2019.04.013>
- Hughes, B. O., & Black, A. J. (1973). The preference of domestic hens for different types of battery cage floor. *British Poultry Science*, *14*, 615–619. <https://doi.org/10.1080/00071667308416071>
- Huskisson, S. M., Jacobson, S. L., Egelkamp, C. L., Ross, S. R., & Hopper, L. M. (2020). Using a touchscreen paradigm to evaluate food preferences and response to novel photographic stimuli of food in three primate species (*Gorilla gorilla gorilla*, *Pan troglodytes*, and *Macaca fuscata*). *International Journal of Primatology*, *41*, 5–23. <https://doi.org/10.1007/s10764-020-00131-0>
- Izzo, G. N., Bashaw, M. J., & Campbell, J. B. (2011). Enrichment and individual differences affect welfare indicators in squirrel monkeys (*Saimiri sciureus*). *Journal of Comparative Psychology*, *125*, 347–352. <https://doi.org/10.1037/a0024294>
- Jennings, M., & Prescott, M. J. (2009). Refinements in husbandry, care and common procedures for non-human primates: Ninth report of the BVAAWF/FRAME/RSPCA/UFAW Joint Working Group on Refinement. *Laboratory Animals*, *43*, 1–47. Retrieved from [10.1258/la.2008.007143](https://doi.org/10.1258/la.2008.007143)
- Johnsen, P. F., Johannesson, T., & Sandøe, P. (2001). Assessment of farm animal welfare at herd level: Many goals, many methods. *Acta Agriculturae Scandinavica, Section A — Animal Science*, *51*, 26–33. <https://doi.org/10.1080/090647001316923027>
- Justice, W. S. M., O'Brien, M. F., Szyszka, O., Shotton, J., Gilmour, J. E. M., Riordan, P., & Wolfensohn, S. (2017). Adaptation of the animal welfare assessment grid (AWAG) for monitoring animal welfare in zoological collections. *Veterinary Record*, *181*, 143–143. <https://doi.org/10.1136/vr.104309>

-
- Juthe, A. (2005). Argument by analogy. *Argumentation*, *19*, 1–27. <https://doi.org/10.1007/s10503-005-2314-9>
- Kagan, R., Carter, S., & Allard, S. (2015). A universal animal welfare framework for zoos. *Journal of Applied Animal Welfare Science*, *18*, S1–S10. <https://doi.org/10.1080/10888705.2015.1075830>
- Kahnau, P., Habedank, A., Diederich, K., & Lewejohann, L. (2020). Behavioral methods for severity assessment. *Animals*, *10*, 1136. <https://doi.org/10.3390/ani10071136>
- Kahnau, P., Jaap, A., Diederich, K., Gygax, L., Rudeck, J., & Lewejohann, L. (2022). Determining the value of preferred goods based on consumer demand in a home-cage based test for mice. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-022-01813-8>
- Keubler, L. M., Hoppe, N., Potschka, H., Talbot, S. R., Vollmar, B., Zechner, D., Häger, C., & Bleich, A. (2020). Where are we heading? Challenges in evidence-based severity assessment. *Laboratory Animals*, *54*, 50–62. <https://doi.org/10.1177/0023677219877216>
- Kikusui, T., Winslow, J. T., & Mori, Y. (2006). Social buffering: Relief from stress and anxiety. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*, 2215–2228. <https://doi.org/10.1098/rstb.2006.1941>
- Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2014). Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *Animals*, *4*, 35–44. <https://doi.org/10.3390/ani4010035>
- King, J. E., & Landau, V. I. (2003). Can chimpanzee (*Pan troglodytes*) happiness be estimated by human raters? *Journal of Research in Personality*, *37*, 1–15.
- Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: A practical introduction* (1. ed). Amsterdam: Elsevier.
- Kirkden, R. D., Edwards, J. S. S., & Broom, D. M. (2003). A theoretical comparison of the consumer surplus and the elasticities of demand as measures of motivational strength. *Animal Behaviour*, *65*, 157–178. <https://doi.org/10.1006/anbe.2002.2035>
- Kirkden, R. D., & Pajor, E. A. (2006). Using preference, motivation and aversion tests to ask scientific questions about animals' feelings. *Applied Animal Behaviour Science*, *100*, 29–47. <https://doi.org/10.1016/j.applanim.2006.04.009>
- Kiros, T. G., Levast, B., Auray, G., Strom, S., van Kessel, J., & Gerdtts, V. (2012). The importance of animal models in the development of vaccines. In S. Baschieri (Ed.), *Innovation in vaccinology* (pp. 251–264). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-4543-8_11
- Knaebe, B., Weiss, C. C., Zimmermann, J., & Hayden, B. Y. (2022). The promise of behavioral tracking systems for advancing primate animal welfare. *Animals*, *12*, 1648. <https://doi.org/10.3390/ani12131648>
- Koolhaas, J. M., Korte, S., De Boer, S., Van Der Vegt, B., Van Reenen, C., Hopster, H., De Jong, I., Ruis, M., & Blokhuis, H. (1999). Coping styles in animals: Current status in behavior and stress-physiology. *Neuroscience and Biobehavioral Reviews*, *23*, 925–935.
- Korte, S. M., Olivier, B., & Koolhaas, J. M. (2007). A new animal welfare concept based on allostasis. *Physiology & Behavior*, *92*, 422–428. <https://doi.org/10.1016/j.physbeh.2006>
-

10.018

- Koster, E. H. W., Crombez, G., Verschuere, B., & De Houwer, J. (2004). Selective attention to threat in the dot probe paradigm: Differentiating vigilance and difficulty to disengage. *Behaviour Research and Therapy*, *42*, 1183–1192. <https://doi.org/10.1016/j.brat.2003.08.001>
- Koster, E. H., Crombez, G., Verschuere, B., Van Damme, S., & Wiersema, J. R. (2006). Components of attentional bias to threat in high trait anxiety: Facilitated engagement, impaired disengagement, and attentional avoidance. *Behaviour Research and Therapy*, *44*, 1757–1771. <https://doi.org/10.1016/j.brat.2005.12.011>
- Krakenberg, V., Siestrup, S., Palme, R., Kaiser, S., Sachser, N., & Richter, S. H. (2020). Effects of different social experiences on emotional state in mice. *Scientific Reports*, *10*, 15255. <https://doi.org/10.1038/s41598-020-71994-9>
- Kremer, L., Klein Holkenborg, S. E. J., Reimert, I., Bolhuis, J. E., & Webb, L. E. (2020). The nuts and bolts of animal emotion. *Neuroscience & Biobehavioral Reviews*, *113*, 273–286. <https://doi.org/10.1016/j.neubiorev.2020.01.028>
- Kret, M. E., Jaasma, L., Bionda, T., & Wijnen, J. G. (2016). Bonobos (*Pan paniscus*) show an attentional bias toward conspecifics' emotions. *Proceedings of the National Academy of Sciences*, *113*, 3761–3766. <https://doi.org/10.1073/pnas.1522060113>
- Kret, M. E., Muramatsu, A., & Matsuzawa, T. (2018). Emotion processing across and within species: A comparison between humans (*Homo sapiens*) and chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, *132*, 395–409. <https://doi.org/10.1037/com0000108>
- Kuraoka, K., & Nakamura, K. (2011). The use of nasal skin temperature measurements in studying emotion in macaque monkeys. *Physiology & Behavior*, *102*, 347–355. <https://doi.org/10.1016/j.physbeh.2010.11.029>
- Kurland, A., & St. Peter, C. C. (2022). Connecting animal trainers and behavior analysts through loopy training. *Journal of the Experimental Analysis of Behavior*. <https://doi.org/10.1002/jeab.779>
- Kurtycz, L. M., Wagner, K. E., & Ross, S. R. (2014). The choice to access outdoor areas affects the behavior of great apes. *Journal of Applied Animal Welfare Science*, *17*, 185–197. <https://doi.org/10.1080/10888705.2014.896213>
- Kuthyar, S., Manus, M. B., & Amato, K. R. (2019). Leveraging non-human primates for exploring the social transmission of microbes. *Current Opinion in Microbiology*, *50*, 8–14. <https://doi.org/10.1016/j.mib.2019.09.001>
- Lagisz, M., Zidar, J., Nakagawa, S., Neville, V., Sorato, E., Paul, E. S., Bateson, M., Mendl, M., & Løvlie, H. (2020). Optimism, pessimism and judgement bias in animals: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *118*, 3–17. <https://doi.org/10.1016/j.neubiorev.2020.07.012>
- Landa, L. (2012). Pain in domestic animals and how to assess it: A review. *Veterinární Medicína*, *57*, 185–192. <https://doi.org/10.17221/5915-VETMED>
- Lattal, K. A. (2010). Delayed reinforcement of operant behavior. *Journal of the Experimental*

- Analysis of Behavior*, 93, 129–139. <https://doi.org/10.1901/jeab.2010.93-129>
- Lautenbacher, S., Huber, C., Schöfer, D., Kunz, M., Parthum, A., Weber, P. G., Roman, C., Griessinger, N., & Sittl, R. (2010). Attentional and emotional mechanisms related to pain as predictors of chronic postoperative pain: A comparison with other psychological and physiological predictors. *Pain*, 151, 722–731. <https://doi.org/10.1016/j.pain.2010.08.041>
- Lea, S. E. (1978). The psychology and economics of demand. *Psychological Bulletin*, 85, 441–466. <https://doi.org/10.1037/0033-2909.85.3.441>
- Leach, M., Thornton, P., & Main, D. (2008). Identification of appropriate measures for the assessment of laboratory mouse welfare. *Animal Welfare*, 10.
- Leaman, J., Latter, J., & Clemence, M. (2014). *Attitudes to animal research in 2014*. Ipsos MORI.
- LeDoux, J. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. New York, NY, US: Simon and Schuster.
- LeDoux, J. (2003). The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology*, 12. <https://doi.org/10.1023/A:1025048802629>
- Lee, T.-H., Sakaki, M., Cheng, R., Velasco, R., & Mather, M. (2014). Emotional arousal amplifies the effects of biased competition in the brain. *Social Cognitive and Affective Neuroscience*, 9, 2067–2077. <https://doi.org/10.1093/scan/nsu015>
- Lee, V. K., Flynt, K. S., Haag, L. M., & Taylor, D. K. (2010). Comparison of the effects of ketamine, ketamine-medetomidine, and ketamine- midazolam on physiologic parameters and anesthesia-induced stress in rhesus (*Macaca mulatta*) and cynomolgus (*Macaca fascicularis*) macaques. *Journal of the American Association for Laboratory Animal Science*, 49, 57–63.
- Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, 63, 1279–1292. <https://doi.org/10.3758/BF03194543>
- Leenaars, C. H. C., Tsaioun, K., Stafleu, F., Rooney, K., Meijboom, F., Ritskes-Hoitinga, M., & Bleich, A. (2020). Reviewing the animal literature: How to describe and choose between different types of literature reviews. *Laboratory Animals*. <https://doi.org/10.1177/0023677220968599>
- Leotti, L. A., Iyengar, S. S., & Ochsner, K. N. (2010). Born to choose: The origins and value of the need for control. *Trends in Cognitive Sciences*, 14, 457–463. <https://doi.org/10.1016/j.tics.2010.08.001>
- Lewejohann, L., Schwabe, K., Häger, C., & Jirkof, P. (2020). Impulse for animal welfare outside the experiment. *Laboratory Animals*, 54, 150–158. <https://doi.org/10.1177/0023677219891754>
- Lieberman, D. A., McIntosh, D. C., & Thomas, G. V. (1979). Learning when reward is delayed: A marking hypothesis. *Journal of Experimental Psychology: Animal Behavior Processes*, 5, 224. <https://doi.org/10.1037/0097-7403.5.3.224>
- Lilly, A. A., Mehlman, P. T., & Higley, J. D. (1999). Trait-like immunological and hematological measures in female rhesus across varied environmental conditions. *American Journal of Primatology*, 48, 197–223. [https://doi.org/10.1002/\(SICI\)1098-2345\(1999\)48:](https://doi.org/10.1002/(SICI)1098-2345(1999)48:)

- 3%3C197::AID-AJP3%3E3.0.CO;2-Y
- Lloyd, M. H., Foden, B. W., & Wolfensohn, S. E. (2008). Refinement: Promoting the three Rs in practice. *Laboratory Animals*, *42*, 284–293. <https://doi.org/10.1258/la.2007.007045>
- Lund, T. B., Lassen, J., & Sandøe, P. (2012). Public attitude formation regarding animal research. *Anthrozoös*, *25*, 475–490. <https://doi.org/10.2752/175303712X13479798785896>
- MacLeod, C., Mathews, A., & Tata, P. (1986). Attentional bias in emotional disorders. *Journal of Abnormal Psychology*, *95*, 15. <https://doi.org/10.1037/0021-843X.95.1.15>
- Macleod, M. R., Ebrahim, S., & Roberts, I. (2005). Surveying the literature from animal experiments: Systematic review and meta-analysis are important contributions. *British Medical Journal*, *331*, 110.3. <https://doi.org/10.1136/bmj.331.7508.110-b>
- Maestriperi, D., Schino, G., Aureli, F., & Troisi, A. (1992). A modest proposal: Displacement activities as an indicator of emotions in primates. *Animal Behaviour*, *44*, 967–979. [https://doi.org/10.1016/S0003-3472\(05\)80592-5](https://doi.org/10.1016/S0003-3472(05)80592-5)
- Magden, E. R., Haller, R. L., Thiele, E. J., Buchl, S. J., Lambeth, S. P., & Schapiro, S. J. (2013). Acupuncture as an adjunct therapy for osteoarthritis in chimpanzees (*Pan troglodytes*). *Journal of the American Association for Laboratory Animal Science*, *52*, 6.
- Magden, E. R., Sleeper, M. M., Buchl, S. J., Jones, R. A., Thiele, E. J., & Wilkerson, G. K. (2016). Use of an implantable loop recorder in a chimpanzee (*Pan troglodytes*) to monitor cardiac arrhythmias and assess the effects of acupuncture and laser therapy. *Comparative Medicine*, *66*, 7.
- Marchant-Forde, J. N. (2015). The science of animal behavior and welfare: Challenges, opportunities, and global perspective. *Frontiers in Veterinary Science*, *2*, 16. <https://doi.org/10.3389/fvets.2015.00016>
- Martin, C. F., Muramatsu, A., & Matsuzawa, T. (2022). Apex and ApeTouch: Development of a portable touchscreen system and software for primates at zoos. *Animals*, *12*, 1660. <https://doi.org/10.3390/ani12131660>
- Mason, G., & Mendl, M. (1993). Why is there no simple way of measuring animal welfare? *Animal Welfare*, *2*, 301–319.
- Mason, S., Premereur, E., Pelekanos, V., Emberton, A., Honess, P., & Mitchell, A. S. (2019). Effective chair training methods for neuroscience research involving rhesus macaques (*Macaca mulatta*). *Journal of Neuroscience Methods*, *317*, 82–93. <https://doi.org/10.1016/j.jneumeth.2019.02.001>
- Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*, 1281–1289. <https://doi.org/10.1038/s41593-018-0209-y>
- Matsuzawa, T., Hasegawa, Y., Gotoh, S., & Wada, K. (1983). One-trial long-lasting food-aversion learning in wild Japanese monkeys (*Macaca fuscata*). *Behavioral and Neural Biology*, *39*, 155–159. [https://doi.org/10.1016/S0163-1047\(83\)90791-4](https://doi.org/10.1016/S0163-1047(83)90791-4)
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: Taylor and Francis, CRC Press.

- McNally, R. J. (2019). Attentional bias for threat: Crisis or opportunity? *Clinical Psychology Review, 69*, 4–13. <https://doi.org/10.1016/j.cpr.2018.05.005>
- Meagher, R. (2019). Is boredom an animal welfare concern? *Animal Welfare, 28*, 21–32. <https://doi.org/10.7120/09627286.28.1.021>
- Meagher, R. K. (2009). Observer ratings: Validity and value as a tool for animal welfare research. *Applied Animal Behaviour Science, 119*, 1–14. <https://doi.org/10.1016/j.applanim.2009.02.026>
- Mellen, J., & Sevenich MacPhee, M. (2001). Philosophy of environmental enrichment: Past, present, and future. *Zoo Biology, 20*, 211–226. <https://doi.org/10.1002/zoo.1021>
- Mellor, D. (2016). Moving beyond the “five freedoms” by updating the “five provisions” and introducing aligned “animal welfare aims.” *Animals, 6*, 59. <https://doi.org/10.3390/ani6100059>
- Mench, J. A. (2003). Assessing animal welfare at the farm and group level: A United States perspective. *Animal Welfare, 12*, 493–503.
- Mendl, M., Burman, O. H. P., Parker, R. M. A., & Paul, E. S. (2009). Cognitive bias as an indicator of animal emotion and welfare: Emerging evidence and underlying mechanisms. *Applied Animal Behaviour Science, 118*, 161–181. <https://doi.org/10.1016/j.applanim.2009.02.023>
- Mendl, M., Burman, O. H. P., & Paul, E. S. (2010). An integrative and functional framework for the study of animal emotion and mood. *Proceedings of the Royal Society B: Biological Sciences, 277*, 2895–2904. <https://doi.org/10.1098/rspb.2010.0303>
- Mendl, M., & Paul, E. S. (2020). Animal affect and decision-making. *Neuroscience and Biobehavioral Reviews, 112*, 144–163. <https://doi.org/10.1016/j.neubiorev.2020.01.025>
- Meyerholz, D. K., Beck, A. P., & Singh, B. (2020). Innovative use of animal models to advance scientific research. *Cell and Tissue Research, 380*, 205–206. <https://doi.org/10.1007/s00441-020-03210-z>
- Millot, S., Cerqueira, M., Castanheira, M. F., Øverli, Ø., Martins, C. I. M., & Oliveira, R. F. (2014). Use of conditioned place preference/avoidance tests to assess affective states in fish. *Applied Animal Behaviour Science, 154*, 104–111. <https://doi.org/10.1016/j.applanim.2014.02.004>
- Moberg, G. (2000). Biological response to stress: Implications for animal welfare. In G. P. Moberg & J. A. Mench (Eds.), *The biology of animal stress: Basic principles and implications for animal welfare* (pp. 1–21). New York: CABI Publishing.
- Mogg, K., & Bradley, B. P. (1998). A cognitive-motivational analysis of anxiety. *Behaviour Research and Therapy, 36*, 809–848. [https://doi.org/10.1016/S0005-7967\(98\)00063-1](https://doi.org/10.1016/S0005-7967(98)00063-1)
- Mogg, K., & Bradley, B. P. (2005). Attentional bias in generalized anxiety disorder versus depressive disorder. *Cognitive Therapy and Research, 29*, 29–45. <https://doi.org/10.1007/s10608-005-1646-y>
- Mogg, K., Bradley, B. P., Hyare, H., & Lee, S. (1998). Selective attention to food-related stimuli in hunger: Are attentional biases specific to emotional and psychopathological states, or are they also found in normal drive states? *Behaviour Research and Therapy, 36*, 227–237.

- [https://doi.org/10.1016/S0005-7967\(97\)00062-4](https://doi.org/10.1016/S0005-7967(97)00062-4)
- Morel, P., Ulbrich, P., & Gail, A. (2017). What makes a reach movement effortful? Physical effort discounting supports common minimization principles in decision making and motor control. *PLoS Biology*, *15*, e2001323. <https://doi.org/10.1371/journal.pbio.2001323>
- Morton, D. B., & Griffiths, P. H. (1985). Guidelines on the recognition of pain, distress and discomfort in experimental animals and an hypothesis for assessment. *Vet Rec*, *116*, 431–6. <https://doi.org/10.1136/vr.116.16.431>
- Mueller-Paul, J., Wilkinson, A., Aust, U., Steurer, M., Hall, G., & Huber, L. (2014). Touchscreen performance and knowledge transfer in the red-footed tortoise (*Chelonoidis carbonaria*). *Behavioural Processes*, *106*, 187–192. <https://doi.org/10.1016/j.beproc.2014.06.003>
- Murray, T., O'Brien, J., Sagiv, N., & Garrido, L. (2021). The role of stimulus-based cues and conceptual information in processing facial expressions of emotion. *Cortex*. <https://doi.org/10.1016/j.cortex.2021.08.007>
- Nannoni, E., Valsami, T., Sardi, L., & Martelli, G. (2014). Tail docking in pigs: A review on its short- and long-term consequences and effectiveness in preventing tail biting. *Italian Journal of Animal Science*, *13*, 3095. <https://doi.org/10.4081/ijas.2014.3095>
- Nettle, D., & Bateson, M. (2012). The evolutionary origins of mood and its disorders. *Current Biology*, *22*, R712–R721. <https://doi.org/10.1016/j.cub.2012.06.020>
- Nguyen, H. A. T., Guo, C., & Homberg, J. R. (2020). Cognitive bias under adverse and rewarding conditions: A systematic review of rodent studies. *Frontiers in Behavioral Neuroscience*, *14*, 14. <https://doi.org/10.3389/fnbeh.2020.00014>
- Noonan, M. P., Sallet, J., Mars, R. B., Neubert, F. X., O'Reilly, J. X., Andersson, J. L., Mitchell, A. S., Bell, A. H., Miller, K. L., & Rushworth, M. F. S. (2014). A Neural Circuit Covarying with Social Hierarchy in Macaques. *PLoS Biology*, *12*, e1001940. <https://doi.org/10.1371/journal.pbio.1001940>
- Novak, M. A. (2003). Self-injurious behavior in rhesus monkeys: New insights into its etiology, physiology, and treatment. *American Journal of Primatology*, *59*, 3–19. <https://doi.org/10.1002/ajp.10063>
- Novak, M. A., Hamel, A. F., Kelly, B. J., Dettmer, A. M., & Meyer, J. S. (2013). Stress, the HPA axis, and nonhuman primate well-being: A review. *Applied Animal Behaviour Science*, *143*, 135–149. <https://doi.org/10.1016/j.applanim.2012.10.012>
- Öhman, A. (1986). Face the beast and fear the face: Animal and social fears as prototypes for evolutionary analyses of emotion. *Psychophysiology*, *23*, 123–145. <https://doi.org/10.1111/j.1469-8986.1986.tb00608.x>
- Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General*, *130*, 466. <https://doi.org/10.1037/0096-3445.130.3.466>
- Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*, 483–522. <https://doi.org/10.1037/0033-295X.108.3.483>

- Orlov, T., Yakovlev, V., Hochstein, S., & Zohary, E. (2000). Macaque monkeys categorize images by their ordinal number. *Nature*, *404*, 77–80. <https://doi.org/10.1038/35003571>
- Owen, M. A., Swaisgood, R. R., Czekala, N. M., & Lindburg, D. G. (2005). Enclosure choice and well-being in giant pandas: Is it all about control? *Zoo Biology*, *24*, 475–481. <https://doi.org/10.1002/zoo.20064>
- Pagliari, F., Addessi, E., Sbaffi, A., Tasselli, M. I., & Delfino, A. (2015). Is it patience or motivation? On motivational confounds in intertemporal choice tasks. *Journal of the Experimental Analysis of Behavior*, *103*, 196–217. <https://doi.org/10.1002/jeab.118>
- Palmer, S., Oppler, S. H., & Graham, M. L. (2022). Behavioral management as a coping strategy for managing stressors in primates: The influence of temperament and species. *Biology*, *11*, 423. <https://doi.org/10.3390/biology11030423>
- Panksepp, J. (2011). Cross-species affective neuroscience decoding of the primal affective experiences of humans and related animals. *PLoS ONE*, *6*, e21236. <https://doi.org/10.1371/journal.pone.0021236>
- Panksepp, J. B., & Lahvis, G. P. (2007). Social reward among juvenile mice. *Genes, Brain and Behavior*, *6*, 661–671. <https://doi.org/10.1111/j.1601-183X.2006.00295.x>
- Parr, L. A. (2011). The evolution of face processing in primates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*, 1764–1777. <https://doi.org/10.1098/rstb.2010.0358>
- Paul, E. S., Browne, W., Mendl, M. T., Caplen, G., Trevarthen, A., Held, S., & Nicol, C. J. (2022). Assessing animal welfare: A triangulation of preference, judgement bias and other candidate welfare indicators. *Animal Behaviour*, *186*, 151–177. <https://doi.org/10.1016/j.anbehav.2022.02.003>
- Paul, E. S., Harding, E. J., & Mendl, M. (2005). Measuring emotional processes in animals: The utility of a cognitive approach. *Neuroscience and Biobehavioral Reviews*, *29*, 469–491. <https://doi.org/10.1016/j.neubiorev.2005.01.002>
- Paul, E. S., Sher, S., Tamietto, M., Winkielman, P., & Mendl, M. T. (2020). Towards a comparative science of emotion: Affect and consciousness in humans and animals. *Neuroscience and Biobehavioral Reviews*, *108*, 749–770. <https://doi.org/10.1016/j.neubiorev.2019.11.014>
- Perel, P., Roberts, I., Sena, E., Wheble, P., Briscoe, C., Sandercock, P., Macleod, M., Mignini, L. E., Jayaram, P., & Khan, K. S. (2007). Comparison of treatment effects between animal experiments and clinical trials: Systematic review. *British Medical Journal*, *334*, 197. <https://doi.org/10.1136/bmj.39048.407928.BE>
- Pérez, C., Lucas, F., & Sclafani, A. (1995). Carbohydrate, fat, and protein condition similar flavor preferences in rats using an oral-delay procedure. *Physiology & Behavior*, *57*, 549–554. [https://doi.org/10.1016/0031-9384\(94\)00366-D](https://doi.org/10.1016/0031-9384(94)00366-D)
- Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience*, *35*, 73–89. <https://doi.org/10.1146/annurev-neuro-062111-150525>
- Pfefferle, D., Plümer, S., Burchardt, L., Treue, S., & Gail, A. (2018). Assessment of stress

- responses in rhesus macaques (*Macaca mulatta*) to daily routine procedures in system neuroscience based on salivary cortisol concentrations. *PLoS ONE*, *13*, e0190190. <https://doi.org/10.1371/journal.pone.0190190>
- Pham, T. M., Hagman, B., Codita, A., Van Loo, P. L. P., Strömmer, L., & Baumans, V. (2010). Housing environment influences the need for pain relief during post-operative recovery in mice. *Physiology & Behavior*, *99*, 663–668. <https://doi.org/10.1016/j.physbeh.2010.01.038>
- Phillips, K. A., Bales, K. L., Capitanio, J. P., Conley, A., Czoty, P. W., Hart, B. A. 't, Hopkins, W. D., Hu, S.-L., Miller, L. A., Nader, M. A., Nathanielsz, P. W., Rogers, J., Shively, C. A., & Voytko, M. L. (2014). Why primate models matter. *American Journal of Primatology*, *76*, 801–827. <https://doi.org/10.1002/ajp.22281>
- Pickard, J. (2013). Review of the assessment of cumulative severity and lifetime experience in non-human primates used in neuroscience research. Animal Procedures Committee. Retrieved from www.gov.uk/government/uploads/system/uploads/attachment_data/file/261687/cs_nhp_review_FINAL_2013_corrected.pdf
- Poirier, C., Bateson, M., Gualtieri, F., Armstrong, E. A., Laws, G. C., Boswell, T., & Smulders, T. V. (2019). Validation of hippocampal biomarkers of cumulative affective experience. *Neuroscience and Biobehavioral Reviews*, *101*, 113–121. <https://doi.org/10.1016/j.neubiorev.2019.03.024>
- Poirier, C., Hamed, S. B., Garcia-Saldivar, P., Kwok, S. C., Meguerditchian, A., Merchant, H., Rogers, J., Wells, S., & Fox, A. S. (2021). Beyond MRI: On the scientific value of combining non-human primate neuroimaging with metadata. *NeuroImage*, *228*, 117679. <https://doi.org/10.1016/j.neuroimage.2020.117679>
- Ponce, C. R., Genecin, M. P., Perez-Melara, G., & Livingstone, M. S. (2016). Automated chair-training of rhesus macaques. *Journal of Neuroscience Methods*, *263*, 75–80. <https://doi.org/10.1016/j.jneumeth.2016.01.024>
- Poole, T. (1997). Happy animals make good science. *Laboratory Animals*, *31*, 116–124. <https://doi.org/10.1258/002367797780600198>
- Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*, 3–25. <https://doi.org/10.1080/00335558008248231>
- Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*, 25–42. <https://doi.org/10.1146/annurev.ne.13.030190.000325>
- Pound, P., Ebrahim, S., Sandercock, P., Bracken, M. B., & Roberts, I. (2004). Where is the evidence that animal research benefits humans? *British Medical Journal*, *328*, 514–517. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC351856/>
- Prescott, M. J., Clark, C., Dowling, W. E., & Shurtleff, A. C. (2021). Opportunities for refinement of non-human primate vaccine studies. *Vaccines*, *9*, 284. <https://doi.org/10.3390/vaccines9030284>
- Prescott, M. J., Langermans, J. A., & Ragan, I. (2017). Applying the 3Rs to non-human primate research: Barriers and solutions. *Drug Discovery Today: Disease Models*, *23*, 51–56. <https://doi.org/10.1016/j.ddmod.2017.11.001>
- Price, R. B., Kuckertz, J. M., Siegle, G. J., Ladouceur, C. D., Silk, J. S., Ryan, N. D., Dahl,

- R. E., & Amir, N. (2015). Empirical recommendations for improving the stability of the dot-probe task in clinical research. *Psychological Assessment, 27*, 365–376. <https://doi.org/10.1037/pas0000036>
- R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Raulo, A., & Dantzer, B. (2018). Associations between glucocorticoids and sociality across a continuum of vertebrate social behavior. *Ecology and Evolution, 8*, 7697–7716. <https://doi.org/10.1002/ece3.4059>
- Rennie, A. E., & Buchanan-Smith, H. M. (2006a). Refinement of the use of non-human primates in scientific research. Part I: The influence of humans. *Animal Welfare, 15*, 203. Retrieved from <http://euprimgvets.eu/uploads/publications/public/RennieBSpt1.pdf>
- Rennie, A., & Buchanan-Smith, H. (2006b). Refinement of the use of non-human primates in scientific research. Part II: Housing, husbandry and acquisition. *Animal Welfare, 25*.
- Rennie, A., & Buchanan-Smith, H. (2006c). Refinement of the use of non-human primates in scientific research. Part III: Refinement of procedures. *Animal Welfare, 23*.
- Revusky, S. (1971). The role of interference in association over a delay. In W. K. Honig & P. H. R. James (Eds.), *Animal memory* (pp. 155–213). Academic Press. <https://doi.org/10.1016/B978-0-12-355050-7.50009-6>
- Riba-Hernández, P., Stoner, K. E., & Lucas, P. W. (2005). Sugar concentration of fruits and their detection via color in the Central American spider monkey (*Ateles geoffroyi*). *American Journal of Primatology, 67*, 411–423. <https://doi.org/10.1002/ajp.20196>
- Richards, R. W. (1981). A comparison of signaled and unsignaled delay of reinforcement. *Journal of the Experimental Analysis of Behavior, 35*, 145–152. <https://doi.org/10.1901/jeab.1981.35-145>
- Richter, S. H., Schick, A., Hoyer, C., Lankisch, K., Gass, P., & Vollmayr, B. (2012). A glass full of optimism: Enrichment effects on cognitive bias in a rat model of depression. *Cognitive, Affective, & Behavioral Neuroscience, 12*, 527–542. <https://doi.org/10.3758/s13415-012-0101-2>
- Ritvo, S. E., & MacDonald, S. E. (2020). Preference for free or forced choice in Sumatran orangutans (*Pongo abelii*). *Journal of the Experimental Analysis of Behavior, 113*, 419–434. <https://doi.org/10.1002/jeab.584>
- Rix, A., Drude, N., Mrugalla, A., Mottaghy, F. M., Tolba, R. H., & Kiessling, F. (2020). Performance of severity parameters to detect chemotherapy-induced pain and distress in mice. *Laboratory Animals, 54*, 452–460. <https://doi.org/10.1177/0023677219883327>
- Roberts, S. J., & Platt, M. L. (2005). Effects of isosexual pair-housing on biomedical implants and study participation in male macaques. *Journal of the American Association for Laboratory Animal Science, 44*, 13–18. Retrieved from <http://www.ingentaconnect.com/content/aalas/jaalas/2005/00000044/00000005/art00002>
- Roberts, W. A., & Mazmanian, D. S. (1988). Concept learning at different levels of abstraction by pigeons, monkeys, and people. *Journal of Experimental Psychology: Animal Behavior Processes, 14*, 247. <https://doi.org/10.1037/0097-7403.14.3.247>

- Robinson, L. M., Altschul, D. M., Wallace, E. K., Úbeda, Y., Llorente, M., Machanda, Z., Slocombe, K. E., Leach, M. C., Waran, N. K., & Weiss, A. (2017). Chimpanzees with positive welfare are happier, extraverted, and emotionally stable. *Applied Animal Behaviour Science*, *191*, 90–97. <https://doi.org/10.1016/j.applanim.2017.02.008>
- Robinson, L. M., Coleman, K., Capitanio, J. P., Gottlieb, D. H., Handel, I. G., Adams, M. J., Leach, M. C., Waran, N. K., & Weiss, A. (2018). Rhesus macaque personality, dominance, behavior, and health. *American Journal of Primatology*, e22739. <https://doi.org/10.1002/ajp.22739>
- Robinson, L. M., Waran, N. K., Handel, I., & Leach, M. C. (2021). Happiness, welfare, and personality in rhesus macaques (*Macaca mulatta*). *Applied Animal Behaviour Science*, *236*, 105268. <https://doi.org/10.1016/j.applanim.2021.105268>
- Robinson, L. M., Waran, N. K., Leach, M. C., Morton, F. B., Paukner, A., Lonsdorf, E., Handel, I., Wilson, V. A., Brosnan, S. F., & Weiss, A. (2016). Happiness is positive welfare in brown capuchins (*Sapajus apella*). *Applied Animal Behaviour Science*, *181*, 145–151. <https://doi.org/10.1016/j.applanim.2016.05.029>
- Roelfsema, P. R., & Treue, S. (2014). Basic neuroscience research with nonhuman primates: A small but indispensable component of biomedical research. *Neuron*, *82*, 1200–1204. <https://doi.org/10.1016/j.neuron.2014.06.003>
- Roelofs, S., Boleij, H., Nordquist, R. E., & van der Staay, F. J. (2016). Making decisions under ambiguity: Judgment bias tasks for assessing emotional state in animals. *Frontiers in Behavioral Neuroscience*, *10*, 119. <https://doi.org/10.3389/fnbeh.2016.00119>
- Romero, L. M., & Beattie, U. K. (2021). Common myths of glucocorticoid function in ecology and conservation. *Journal of Experimental Zoology Part A: Ecological and Integrative Physiology*, *337*, 7–14. <https://doi.org/10.1002/jez.2459>
- Rosati, A. G., Stevens, J. R., Hare, B., & Hauser, M. D. (2007). The evolutionary origins of human patience: Temporal preferences in chimpanzees, bonobos, and human adults. *Current Biology*, *17*, 1663–1668. <https://doi.org/10.1016/j.cub.2007.08.033>
- Roughan, J. V., Coulter, C. A., Flecknell, P. A., Thomas, H. D., & Sufka, K. J. (2014). The conditioned place preference test for assessing welfare consequences and potential refinements in a mouse bladder cancer model. *PLoS ONE*, *9*, e103362. <https://doi.org/10.1371/journal.pone.0103362>
- Rudaizky, D., Basanovic, J., & MacLeod, C. (2014). Biased attentional engagement with, and disengagement from, negative information: Independent cognitive pathways to anxiety vulnerability? *Cognition and Emotion*, *28*, 245–259. <https://doi.org/10.1080/02699931.2013.815154>
- Rushen, J. (2000). Some issues in the interpretation of behavioural responses to stress. In M. GP & M. JA (Eds.), *The biology of animal stress: Basic principles and implications for animal welfare* (pp. 23–41). New York: CABI Publishing.
- Russell, W. M., & Burch, R. L. (1959). *The principles of humane experimental technique*. London, UK: Methuen.
- Rygula, R., Abumaria, N., Flügge, G., Fuchs, E., Rüter, E., & Havemann-Reinecke, U. (2005).

- Anhedonia and motivational deficits in rats: Impact of chronic social stress. *Behavioural Brain Research*, *162*, 127–134. <https://doi.org/10.1016/j.bbr.2005.03.009>
- Safarjan, W. R., & D'Amato, M. R. (1981). One-trial, long-delay, conditioned preference in rats. *The Psychological Record*, *31*, 413–426. <https://doi.org/10.1007/BF03394753>
- Sallet, J., Mars, R. B., Noonan, M. P., Andersson, J. L., O'Reilly, J. X., Jbabdi, S., Crosson, P. L., Jenkinson, M., Miller, K. L., & Rushworth, M. F. S. (2011). Social network size affects neural circuits in macaques. *Science*, *334*, 697–700. <https://doi.org/10.1126/science.1210027>
- Salvador-Oliván, J. A., Marco-Cuenca, G., & Arquero-Avilés, R. (2019). Errors in search strategies used in systematic reviews and their effects on information retrieval. *Journal of the Medical Library Association : JMLA*, *107*, 210–221. <https://doi.org/10.5195/jmla.2019.567>
- Sambrook, T. D., & Buchanan-Smith, H. M. (1997). Control and complexity in novel object enrichment. *Animal Welfare*, *6*, 207–216.
- Sánchez-Solano, K. G., Morales-Mávil, J. É., Laska, M., Melin, A., & Hernández-Salazar, L. T. (2020). Visual detection and fruit selection by the mantled howler monkey (*Alouatta palliata*). *American Journal of Primatology*, *82*. <https://doi.org/10.1002/ajp.23186>
- Sato, Y., Sakai, Y., & Hirata, S. (2021). Computerized intertemporal choice task in chimpanzees (*Pan troglodytes*) with/without postreward delay. *Journal of Comparative Psychology*, *135*, 185–195. <https://doi.org/10.1037/com0000254>
- Schapiro, S. J., & Lambeth, S. P. (2007). Control, choice, and assessments of the value of behavioral management to nonhuman primates in captivity. *Journal of Applied Animal Welfare Science*, *10*, 39–47. <https://doi.org/10.1080/10888700701277345>
- Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, *1*, 103–113. <https://doi.org/10.1111/j.2041-210X.2010.00012.x>
- Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*, 416–420. <https://doi.org/10.1093/beheco/arn145>
- Schmitt, V., & Fischer, J. (2011). Representational format determines numerical competence in monkeys. *Nature Communications*, *2*, 257. <https://doi.org/10.1038/ncomms1262>
- Schmitt, V., Schloegl, C., & Fischer, J. (2014). Seeing the experimenter influences the response to pointing cues in long-tailed macaques. *PLoS ONE*, *9*, e91348. <https://doi.org/10.1371/journal.pone.0091348>
- Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality*, *19*, 595–605. <https://doi.org/10.1002/per.554>
- Service, A. and P. H. I. (2019). Annual report animal usage by fiscal year (2019): Total number of animals research facilities used for regulated activities. United States Department of Agriculture. Retrieved from https://www.aphis.usda.gov/animal_welfare/downloads/reports/fy19-summary-report-column-F.pdf
- Shechner, T., Pelc, T., Pine, D. S., Fox, N. A., & Bar-Haim, Y. (2012). Flexible attention deployment in threatening contexts: An instructed fear conditioning study. *Emotion*, *12*, 1041–1049. <https://doi.org/10.1037/a0027072>

- Sheriff, M. J., Dantzer, B., Delehanty, B., Palme, R., & Boonstra, R. (2011). Measuring stress in wildlife: Techniques for quantifying glucocorticoids. *Oecologia*, *166*, 869–887.
- Sherwen, S., Hemsworth, L., Beausoleil, N., Embury, A., & Mellor, D. (2018). An Animal Welfare Risk Assessment Process for Zoos. *Animals*, *8*, 130. <https://doi.org/10.3390/ani8080130>
- Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, *6*, 1317–1322. <https://doi.org/10.1038/nn1150>
- Sibal, L. R., & Samson, K. J. (2001). Nonhuman primates: A critical role in current disease research. *ILAR Journal*, *42*, 74–84. Retrieved from <http://ilarjournal.oxfordjournals.org/content/42/2/74.short>
- Simpson, E. A., Robinson, L. M., & Paukner, A. (2019). Infant rhesus macaque (*Macaca mulatta*) personality and subjective well-being. *PLoS ONE*, *14*, e0226747. <https://doi.org/10.1371/journal.pone.0226747>
- Sipos, M. L., Bar-Haim, Y., Abend, R., Adler, A. B., & Bliese, P. D. (2014). Postdeployment threat-related attention bias interacts with combat exposure to account for PTSD and anxiety symptoms in soldiers. *Depression and Anxiety*, *31*, 124–129. <https://doi.org/10.1002/da.22157>
- Sloan Wilson, D., Clark, A. B., Coleman, K., & Dearstyne, T. (1994). Shyness and boldness in humans and other animals. *Trends in Ecology & Evolution*, *9*, 442–446. [https://doi.org/10.1016/0169-5347\(94\)90134-1](https://doi.org/10.1016/0169-5347(94)90134-1)
- Sneddon, L. U. (2017). Pain in laboratory animals: A possible confounding factor? *Alternatives to Laboratory Animals*, *45*, 161–164. <https://doi.org/10.1177/026119291704500309>
- Sormaz, M., Young, A. W., & Andrews, T. J. (2016). Contributions of feature shapes and surface cues to the recognition of facial expressions. *Vision Research*, *127*, 1–10. <https://doi.org/10.1016/j.visres.2016.07.002>
- Špinková, M., & Wemelsfelder, F. (2011). Environmental Challenge and Animal Agency. *Animal Welfare*, *2*, 27–44. <https://doi.org/10.1079/9781845936594.0027>
- Stansfield, C., O'Mara-Eves, A., & Thomas, J. (2017). Text mining for search term development in systematic reviewing: A discussion of some methods and challenges. *Research Synthesis Methods*, *8*, 355–365. <https://doi.org/10.1002/jrsm.1250>
- Staugaard, S. R. (2009). Reliability of two versions of the dot-probe task using photographic faces. *Psychology Science Quarterly*, *51*, 339–350.
- Stevens, J. R., Hallinan, E. V., & Hauser, M. D. (2005). The ecology and evolution of patience in two New World monkeys. *Biology Letters*, *1*, 223–226. <https://doi.org/10.1098/rsbl.2004.0285>
- Suzuki, S., & Matsuzawa, T. (1997). Choice between two discrimination tasks in chimpanzees (*Pan troglodytes*). *Japanese Psychological Research*, *39*, 226–235. <https://doi.org/10.1111/1468-5884.00056>
- Taira, K., & Rolls, E. T. (1996). Receiving grooming as a reinforcer for the monkey. *Physiology & Behavior*, *59*, 1189–1192. [https://doi.org/10.1016/0031-9384\(95\)02213-9](https://doi.org/10.1016/0031-9384(95)02213-9)
- Tardif, S. D., Coleman, K., Hobbs, T. R., & Lutz, C. (2013). IACUC review of nonhuman

- primate research. *ILAR Journal*, *54*, 234–245. <https://doi.org/10.1093/ilar/ilt040>
- Temple, D., Manteca, X., Dalmau, A., & Velarde, A. (2013). Assessment of test–retest reliability of animal-based measures on growing pig farms. *Livestock Science*, *151*, 35–45. <https://doi.org/10.1016/j.livsci.2012.10.012>
- Testard, C., Brent, L. J. N., Andersson, J., Chiou, K. L., Negron-Del Valle, J. E., DeCasien, A. R., ... Sallet, J. (2022). Social connections predict brain structure in a multidimensional free-ranging primate society. *Science Advances*, *8*, eabl5794. <https://doi.org/10.1126/sciadv.abl5794>
- Thomas, G. V., Lieberman, D. A., McIntosh, D. C., & Ronaldson, P. (1983). The role of marking when reward is delayed. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*, 401–411. <https://doi.org/10.1037/0097-7403.9.4.401>
- Treue, S., & Lemon, R. (2022). The indispensable contribution of nonhuman primates to biomedical research. In L. M. Robinson & A. Weiss (Eds.), *Nonhuman Primate Welfare: From History, Science, and Ethics to Practice*. Cham, Switzerland: Springer.
- Trevarthen, A. C., Kappel, S., Roberts, C., Finnegan, E. M., Paul, E. S., Planas-Sitjà, I., Mendl, M. T., & Fureix, C. (2019). Measuring affect-related cognitive bias: Do mice in opposite affective states react differently to negative and positive stimuli? *PLoS ONE*, *14*, e0226438. <https://doi.org/10.1371/journal.pone.0226438>
- Trimmer, P., Paul, E., Mendl, M., McNamara, J., & Houston, A. (2013). On the evolution and optimality of mood states. *Behavioral Sciences*, *3*, 501–521. <https://doi.org/10.3390/bs3030501>
- Troisi, A. (2002). Displacement activities as a behavioral measure of stress in nonhuman primates and human subjects. *Stress*, *5*, 47–54. <https://doi.org/10.1080/102538902900012378>
- Truelove, M. A., Martin, J. E., Langford, F. M., & Leach, M. C. (2020). The identification of effective welfare indicators for laboratory-housed macaques using a Delphi consultation process. *Scientific Reports*, *10*, 20402. <https://doi.org/10.1038/s41598-020-77437-9>
- Turner, P. V. (2020). Moving beyond the absence of pain and distress: Focusing on positive animal welfare. *ILAR Journal*, *60*, 366–372. <https://doi.org/10.1093/ilar/ilaa017>
- Tuytens, F. A. M., de Graaf, S., Heerkens, J. L. T., Jacobs, L., Nalon, E., Ott, S., Stadig, L., Van Laer, E., & Ampe, B. (2014). Observer bias in animal behaviour research: Can we believe what we score, if we score what we believe? *Animal Behaviour*, *90*, 273–280. <https://doi.org/10.1016/j.anbehav.2014.02.007>
- Tzschentke, T. M. (1998). Measuring reward with the conditioned place preference paradigm: A comprehensive review of drug effects, recent progress and new issues. *Progress in Neurobiology*, *56*, 613–672. [https://doi.org/10.1016/S0301-0082\(98\)00060-4](https://doi.org/10.1016/S0301-0082(98)00060-4)
- Tzschentke, T. M. (2007). Measuring reward with the conditioned place preference (CPP) paradigm: Update of the last decade. *Addiction Biology*, *12*, 227–462. <https://doi.org/10.1111/j.1369-1600.2007.00070.x>
- Ullman-Culler, M. H. (1999). Body condition scoring: A rapid and accurate method for assessing health status in mice. *Laboratory Animal Science*, *49*, 5.
- Ullmann, K., Jourdan, T., Kock, M., Unger, J., Schulz, A., Thöne-Reineke, C., & Abramjuk, C.

- (2018). Recommendations for the development and use of Score Sheets as a tool for applied refinement. *Berliner und Münchener Tierärztliche Wochenschrift*, *131*, 292–298. Retrieved from <https://www.cabdirect.org/cabdirect/abstract/20183235029>
- Unakafov, A. M., Möller, S., Kagan, I., Gail, A., Treue, S., & Wolf, F. (2018). Using imaging photoplethysmography for heart rate estimation in non-human primates. *PLoS ONE*, *13*, e0202581. <https://doi.org/10.1371/journal.pone.0202581>
- van der Mierden, S., Hooijmans, C. R., Tillema, A. H., Rehn, S., Bleich, A., & Leenaars, C. H. (2022). Laboratory animals search filter for different literature databases: PubMed, Embase, Web of Science and PsycINFO. *Laboratory Animals*, *56*, 279–286. <https://doi.org/10.1177/00236772211045485>
- van Rooijen, R., Ploeger, A., & Kret, M. E. (2017). The dot-probe task to measure emotional attention: A suitable measure in comparative studies? *Psychonomic Bulletin & Review*, *24*, 1686–1717. <https://doi.org/10.3758/s13423-016-1224-1>
- Veasey, J. S. (2020a). Assessing the psychological priorities for optimising captive Asian elephant (*Elephas maximus*) welfare. *Animals*, *10*, 39. <https://doi.org/10.3390/ani10010039>
- Veasey, J. S. (2020b). Can zoos ever be big enough for large wild animals? A review using an expert panel assessment of the psychological priorities of the Amur tiger (*Panthera tigris altaica*) as a model species. *Animals*, *10*, 1536. <https://doi.org/10.3390/ani10091536>
- Veerapa, E., Grandgenevre, P., El Fayoumi, M., Vinnac, B., Haelewyn, O., Szaffarczyk, S., Vaiva, G., & D'Hondt, F. (2020). Attentional bias towards negative stimuli in healthy individuals and the effects of trait anxiety. *Scientific Reports*, *10*, 11826. <https://doi.org/10.1038/s41598-020-68490-5>
- Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P., Paananen, T., & Gelman, A. (2020). Package loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models (Version R package version 2.4.1). R package version 2.4.1. Retrieved from <https://mc-stan.org/loo/>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 1: Behavioural study. *European Journal of Neuroscience*, *11*, 1223–1238. <https://doi.org/10.1046/j.1460-9568.1999.00530.x>
- von Borell, E., Langbein, J., Després, G., Hansen, S., Leterrier, C., Marchant-Forde, J., Marchant-Forde, R., Minero, M., Mohr, E., Prunier, A., Valance, D., & Veissier, I. (2007). Heart rate variability as a measure of autonomic regulation of cardiac activity for assessing stress and welfare in farm animals — A review. *Physiology & Behavior*, *92*, 293–316. <https://doi.org/10.1016/j.physbeh.2007.01.007>
- Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
- Waechter, S., Nelson, A. L., Wright, C., Hyatt, A., & Oakman, J. (2014). Measuring attentional bias to threat: Reliability of dot probe and eye movement indices. *Cognitive Therapy and*

- Research*, 38, 313–333. <https://doi.org/10.1007/s10608-013-9588-2>
- Wald, I., Shechner, T., Bitton, S., Holoshitz, Y., Charney, D. S., Muller, D., Fox, N. A., Pine, D. S., & Bar-Haim, Y. (2011). Attention bias away from threat during life threatening danger predicts PTSD symptoms at one-year follow-up. *Depression and Anxiety*, 28, 406–411. <https://doi.org/10.1002/da.20808>
- Wallis, L. J., Range, F., Kubinyi, E., Chapagain, D., Serra, J., & Huber, L. (2017). Utilising dog-computer interactions to provide mental stimulation in dogs especially during ageing. In *Proceedings of the Fourth International Conference on Animal-Computer Interaction* (pp. 1–12). Milton Keynes, UK: ACM. <https://doi.org/10.1145/3152130.3152146>
- Walton, M. E., Kennerley, S. W., Bannerman, D. M., Phillips, P. E. M., & Rushworth, M. F. S. (2006). Weighing up the benefits of work: Behavioral and neural analyses of effort-related decision making. *Neural Networks*, 19, 1302–1314. <https://doi.org/10.1016/j.neunet.2006.03.005>
- Webb, L. E., Veenhoven, R., Harfeld, J. L., & Jensen, M. B. (2019). What is animal happiness? *Annals of the New York Academy of Sciences*, 1438, 62–76. <https://doi.org/10.1111/nyas.13983>
- Webb, S. J. N., Hau, J., & Schapiro, S. J. (2018). Refinements to captive chimpanzee (*Pan troglodytes*) care: A self-medication paradigm. *Animal Welfare*, 27, 327–341. <https://doi.org/10.7120/09627286.27.4.327>
- Webster, J. (1994). Assessment of animal welfare: The five freedoms. In *Animal welfare: A cool eye towards eden* (pp. 10–14). Oxford, UK: Blackwell Science.
- Wegener, D., Oh (□□□), D. Q. P., Lukaß, H., Böer, M., & Kreiter, A. K. (2021). Blood Analysis of Laboratory Macaca mulatta Used for Neuroscience Research: Investigation of Long-Term and Cumulative Effects of Implants, Fluid Control, and Laboratory Procedures. *eNeuro*, 8. <https://doi.org/10.1523/ENEURO.0284-21.2021>
- Weiss, A., Adams, M. J., & King, J. E. (2011a). Happy orang-utans live longer lives. *Biology Letters*, 7, 872–874. <https://doi.org/10.1098/rsbl.2011.0543>
- Weiss, A., Adams, M. J., Widdig, A., & Gerald, M. S. (2011b). Rhesus macaques (*Macaca mulatta*) as living fossils of hominoid personality and subjective well-being. *Journal of Comparative Psychology*, 125, 72–83. <https://doi.org/10.1037/a0021187>
- Welfare Quality Network. (2022). Assessment protocols. Retrieved from <http://www.welfarequalitynetwork.net/en-us/reports/assessment-protocols/>
- Wemelsfelder, F. (2007). How animals communicate quality of life: The qualitative assessment of behaviour. *WBI Studies Repository*. Retrieved from http://animalstudiesrepository.org/acwp_asie/100/
- Wemelsfelder, F., Hunter, E. A., Mendl, M. T., & Lawrence, A. B. (2000). The spontaneous qualitative assessment of behavioural expressions in pigs: First explorations of a novel methodology for integrative animal welfare measurement. *Applied Animal Behaviour Science*, 67, 193–215. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0168159199000933>
- Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach.

- Cognitive, Affective, & Behavioral Neuroscience*, 15, 395–415. <https://doi.org/10.3758/s13415-015-0334-y>
- Whytt, H. R., Main, D. C. J., Greent, L. E., & Webster, A. J. F. (2003). Animal-based measures for the assessment of welfare state of dairy cattle, pigs and laying hens: Consensus of expert opinion. *Animal Welfare*, 12, 205–217.
- Whitham, J. C., & Wielebnowski, N. (2009). Animal-based welfare monitoring: Using keeper ratings as an assessment tool. *Zoo Biology*, 28, 545–560. <https://doi.org/10.1002/zoo.20281>
- Whitham, J. C., & Wielebnowski, N. (2013). New directions for zoo animal welfare science. *Applied Animal Behaviour Science*, 147, 247–260. <https://doi.org/10.1016/j.applanim.2013.02.004>
- Whittaker, A. L., Golder-Dewar, B., Triggs, J. L., Sherwen, S. L., & McLelland, D. J. (2021). Identification of animal-based welfare indicators in captive reptiles: A Delphi consultation survey. *Animals*, 11, 2010. <https://doi.org/10.3390/ani11072010>
- Whitten, P. L., Stavisky, R., Aureli, F., & Russell, E. (1998). Response of fecal cortisol to stress in captive chimpanzees (*Pan troglodytes*). *American Journal of Primatology*, 44, 57–69. [https://doi.org/10.1002/\(SICI\)1098-2345\(1998\)44:1%3C57::AID-AJP5%3E3.0.CO;2-W](https://doi.org/10.1002/(SICI)1098-2345(1998)44:1%3C57::AID-AJP5%3E3.0.CO;2-W)
- Wilson, V., Guenther, A., Øverli, Ø., Seltmann, M. W., & Altschul, D. (2019). Future directions for personality research: Contributing new insights to the understanding of animal behavior. *Animals*, 9, 240. <https://doi.org/10.3390/ani9050240>
- Winters, S., Dubuc, C., & Higham, J. P. (2015). Perspectives: The looking time experimental paradigm in studies of animal visual perception and cognition. *Ethology*, 121, 625–640. <https://doi.org/10.1111/eth.12378>
- Wolfensohn, S., Sharpe, S., Hall, I., Lawrence, S., Kitchen, S., & Dennis, M. (2015). Refinement of welfare through development of a quantitative system for assessment of lifetime experience. *Animal Welfare*, 24, 139–149. <https://doi.org/10.7120/09627286.24.2.139>
- Wolfensohn, S., Shotton, J., Bowley, H., Davies, S., Thompson, S., & Justice, W. (2018). Assessment of welfare in zoo animals: Towards optimum quality of life. *Animals*, 8, 110. <https://doi.org/10.3390/ani8070110>
- Würbel, H. (2001). Ideal homes? Housing effects on rodent brain and behaviour. *Trends in Neurosciences*, 24, 207–211. [https://doi.org/10.1016/S0166-2236\(00\)01718-5](https://doi.org/10.1016/S0166-2236(00)01718-5)

Chapter 8

Appendices

8.1 Appendix A: Supplementary material for Chapter 2 (Choice-based Severity Scale)

8.1.1 Guidelines for Choice-based Severity Assessments in animals

First, it is essential that the animals know what the conditions are and are associated with representative stimuli properly. In our study, we initially trained the monkeys to associate visual stimuli with a small motivational reward from the cage condition cognitive testing system (cage condition) or once seated in the non-human primate chair (lab condition). During training, each condition stimulus was paired with an unrewarded stimulus for several trials each day. Training was repeated until the monkeys learned to select the condition stimuli at least 75 % of trials per trial type in multiple sessions.

Second, it is important to ensure that the experimental setup where the animals are given choices minimizes arousal and influence from the experimenter (Schmitt et al., 2014), which could confound animals' choices, and is unbiased. By placing the experimental setup close to the monkeys' home cage, we minimized arousal by eliminating the need for additional transport (Habedank et al., 2018; Kahnau et al., 2022). Additionally, this setup took advantage of the existing infrastructure and did not require a separate room for testing. We did, however, need to conduct several pilot experiments to reduce environmental bias (see Table 8.2). Additionally, we tested two compartment positions of the cage condition (position on the upper and lower quadrants of the compartment) to control that choices were made for the condition rather than the monkeys' natural tendency to go up (Clarence et al., 2006).

Third, an animals' prior experience with the conditions provided may have an influence on their choice behavior (Arnold & Estep, 1994; Fraser & Nicol, 2011; Habedank et al., 2018). Preferences are difficult to interpret if one option is more familiar than the other as they could be due to novelty seeking or risk aversion (Dawkins, 1977; Habedank et al., 2021). Ideally, animals would have equal experience with all conditions, but such a goal is often impractical in research settings. Moreover, the animals experience with different conditions is arguably what shapes and renders their subjective perspective. Therefore, we emphasize the importance of sufficient general and recent familiarity before Choice-based Severity Assessments. For example, our monkeys were extensively trained in both the cage and lab conditions prior to our study and

the CSS protocol ensured that their most recent experiences were balanced between the two conditions. Importantly, none of the monkeys had chronic implants at the time of testing and only one animal had such an experience previously. Chronic implants require regular inspection and treatment to prevent the development of issues, such as infection, which can cause pain. Therefore, a monkey with these past aversive experiences in association to the lab condition may be less willing—and have a stronger aversion—to participate in related procedures despite no implant treatment taking place.

8.1.2 Supplementary information

8.1.2.1 Basic experimental task training for the Choice-based Severity Assessment

Prior to conducting the Choice-based Severity Assessment, we trained all monkeys to perform a basic experimental task commonly used as a first training step in neuroscience labs (i.e., touch-hold-release). The monkeys learned to perform either a stimulus color (monkey D) or random dot-pattern direction change (monkeys H and E) task, that could be carried out by holding and releasing a sensor in either the cage or lab condition. Both basic experimental tasks are described in detail in the next sections. During the training phase, the monkeys were rewarded with 0.3 ml (one drop) of grape juice for each correct trial (i.e., detecting a change in the stimulus). Prior to any choice testing, we ensured that the monkeys could perform the basic experimental task successfully over 80 % of trials within a session. Training sessions ran for two hours in both cage and lab conditions. From this training data, we examined the level of engagement (i.e., when trials were being conducted) across the two hours of training for multiple training sessions in each condition, as the monkeys often conducted most of the total number of trials in a session within the first hour of training. Then we determined an inter-trial interval for each monkey that would generally indicate that the monkey had completed at least 80 % of the total number trials of these 2-hour training sessions in both conditions (monkey H: 12 min; monkey D: 9 min; monkey E: 8 min). We used this inter-trial interval as an indicator that the monkey was not interested in engaging in the basic experimental task further. This conclusion criterion was important to apply as the monkeys may have viewed excess time spent in either condition as a punishment. The basic experimental task was programmed to automatically stop once the conclusion criteria for the individual had been met, and the experimenter returned the monkeys to their home cage soon afterwards.

8.1.2.1.1 Stimulus color change task In the stimulus color change task (monkey D), a green square (0.2 cm) was presented in the center of the screen for 2.5 s or until the monkey touched the sensor. If the sensor was not engaged, the stimulus disappeared and then reappeared again after 1.4 s. If the monkey engaged the sensor while no stimulus was present, the task emitted a soft beep until the monkey removed his hand. When the sensor was touched when the stimulus was present, the luminance of the square decreased and remained on the screen for a randomly

selected duration between 0.4 and 2 s, as long as the monkey continued to hold the sensor. An ‘error’ sound occurred if the sensor was released early (i.e., early release), followed by a 1 s timeout. Once the randomly selected duration expired, then the square stimulus turned red and the monkey had to release the sensor within 2.5 s to receive a fluid reward and hear a ‘ding’ sound, indicating a correct trial. If the monkey continued to hold the sensor longer than the response window, then the stimulus disappeared and a ‘beep’ sound occurred every 1 s until the monkey released the sensor (i.e., late response). The basic steps of the stimulus color change task are depicted in steps two to four of Figure 2.3 of Chapter 2.

8.1.2.1.2 Random dot-pattern direction change task The basis of the random dot-pattern direction change task (monkeys H and E) was similar the stimulus color change task in that the monkeys was required to respond to a change in a random dot pattern stimulus via a sensor. In this task, a soft ‘beep’ sound occurred every 2.5 s until the monkey initiated a trial by touching the sensor. Immediately after the monkey touched the sensor, a random dot pattern stimulus appeared in the center of the screen (dot size: 0.01 cm in diameter; dot density: 3.54 dots per cm²; speed: 5 cm per s) moving in one direction (of 360 degrees) for a randomly chosen duration between 0.3 and 2.5 s (monkey H: 0.3 to 2.5 s; monkey E: 0.3 to 1.2 s). After the randomly chosen duration expired, the direction of the stimulus changed by 45 degrees clockwise or counterclockwise randomly and the monkey had to respond to the change by releasing the sensor within 0.7 s. An ‘error’ sound and a 1 s timeout occurred if the monkey released the sensor early. If the monkey continued to hold the sensor past the expiration of the response window, then a ‘beep’ sound occurred every 2 s until the monkey released the sensor (i.e., late response). The basic steps of the random dot-pattern direction change task are depicted in steps two to four of Figure 2.3 of Chapter 2.

8.1.2.2 Choice task and condition stimuli training for the Choice-based Severity Assessment

In addition to learning the basic experimental task, the monkeys were trained to engage with the choice task and condition stimuli that were used in the Choice-based Severity Assessment. Generally, the choice task was structured so that the monkey needed to initiate a trial by touching a small square (i.e., start button) appearing in a black box within a colored bar spanning the lower third of the touchscreen of the neutral cognitive testing system. The monkey needed to touch the start button within 5 s for the condition stimuli to be presented, ensuring his attention was directed towards the touchscreen, otherwise the start button disappeared.

Condition stimuli were introduced in two stages. During the familiarization stage, each condition stimulus appeared on the left or right of the touchscreen on its own for 3 s or until touched by the monkey (a 1 to 10 s timeout and ‘error’ sound occurred if stimulus was not touched). Each condition stimulus constituted a 6.6 cm by 6.6 cm picture (lab condition: non-human primate chair; cage condition: cognitive testing system with the cage

condition stimulus on the touchscreen) and a 11 cm by 11 cm colored background specific for the condition (colors differed between monkeys). Once the monkeys touched the presented condition stimulus, a ‘ding’ sound was produced and the stimulus deluminated for 0.75 s before disappearing. The sliding panel of the corresponding quadrant in the testing compartment was opened and closed by the experimenter once the monkey had entered. If the cage condition stimulus had been touched, the same condition stimulus appeared on the cage condition cognitive testing system, which the monkey triggered to disappear (by either touching the touchscreen or using the proximity sensor) to dispense a small motivational reward (i.e., 2 ml bolus of water). If the lab condition stimulus had been touched, the experimenter opened the cage door to the non-human primate chair so that the monkey could climb in and receive the 2 ml bolus once seated. The lab condition stimulus was mounted on a larger (same) colored background to the front of the non-human primate chair so that the monkey could see the stimulus as he entered the chair. After each familiarization training trial, the monkey was returned to the quadrant with the neutral cognitive testing system. The procedure for the differentiation stage of condition stimulus training was the same except that a stimulus associated with a short timeout (i.e., timeout stimulus) was presented simultaneously on the opposite side of the touchscreen. The timeout stimulus was introduced to train the monkey to make an informed choice. If the timeout stimulus was touched, both stimuli disappeared, and an ‘error’ sound was produced. Otherwise if the condition stimulus was touched, the timeout stimulus disappeared immediately and the condition stimulus deluminated for 0.75 s before disappearing.

In both stages, the position and type of stimulus was counterbalanced every four trials. Training sessions generally consisted of 24 trials, counterbalanced by condition stimulus type. Training ended after the monkeys chose the condition stimulus over the timeout stimulus for at least 75 % of trials for each trial type in a session.

8.1.2.3 Fluid preference test

Prior to the third phase, we tested the relative fluid reward preferences of each monkey. We presented five different types of fluid rewards simultaneously once a day for 10 min to each monkey, counterbalancing the position of each option across five days. Monkey E was tested using different fluid rewards than monkeys D and H due to dietary restrictions. We used the highest mean amount (ml) of fluid consumed as the monkey’s preferred fluid reward (monkeys H and E: grape juice; monkey D: banana juice).

8.1.2.4 Preparation of the Choice-based Severity Scale test

During the Choice-based Severity Scale test, the monkeys were given a choice between two tasks differing in the duration of how long they had to hold a proximity sensor (i.e., ‘sensor’),

until they detected a change in the task stimulus (i.e., short and long hold tasks; see Fig. 5 in the main text). Short and long tasks were differentiated by colored stimuli of the same shape and adapted from the basic experimental task the monkey engaged with during the Choice-based Severity Assessment (see ‘Basic experimental task training’).

During training, we introduced the concept of making a choice between colored stimuli by repeatedly presenting one task stimulus with a timeout stimulus (that differed in color) simultaneously, with the position of the stimuli counterbalanced across trials. Over the course of training, we slowly increased the duration of the hold of each task until it reached its corresponding hold range. Specifically, monkey D was trained to hold the short hold task for a randomized duration between 1 and 4 s, whereas the hold for the long hold task was between 10 s and 40 s. Monkeys H and E were trained on shorter hold durations due to time constraints, but these durations still differed in their range by a factor of 10 (short hold: 1 to 3 s; long hold: 10 to 30 s). We added abort penalty so that the monkeys were encouraged to complete trials of both tasks and receive the corresponding fluid reward. The duration of the abort penalty corresponded to the average hold duration (short hold: 2 to 2.5 s; long hold: 20 to 25 s). We provided more reward (of the monkeys’ preferred juice) for long hold trials during training to maintain the interest, compensate for the additional hold duration, and to keep the reward per second roughly equivalent between the two tasks. At the end of training, the difference in reward between the two tasks was 2 or 3 ml, with a payout of approximately 0.11 to 0.13 ml per s of hold (short hold: 0.3 ml for monkey D, 0.2 ml for monkeys H and E; long hold: 3.3 ml for monkey D, 2.2 ml for monkeys H and E). Training sessions were two hours in length, where each task was presented with the timeout stimulus for approximately one hour each (order alternated each session). We checked that the monkeys could perform both tasks successfully (over 80 % correct for each task within a session) prior to testing.

8.1.3 Supplementary results

8.1.3.1 Results of the Choice-based Severity Assessment for the third monkey (E)

We offered a third monkey (E) a choice between performing a basic experimental task in the lower cage or lab condition. Monkey E was only tested on the third phase, where the type of reward was different between the two conditions (cage: water; lab: grape juice). Like monkey D, monkey E exclusively chose the cage condition (100 % of six choice sessions) despite the amount and type of reward per trial being largely in favor of the lab condition (Figure 8.1).

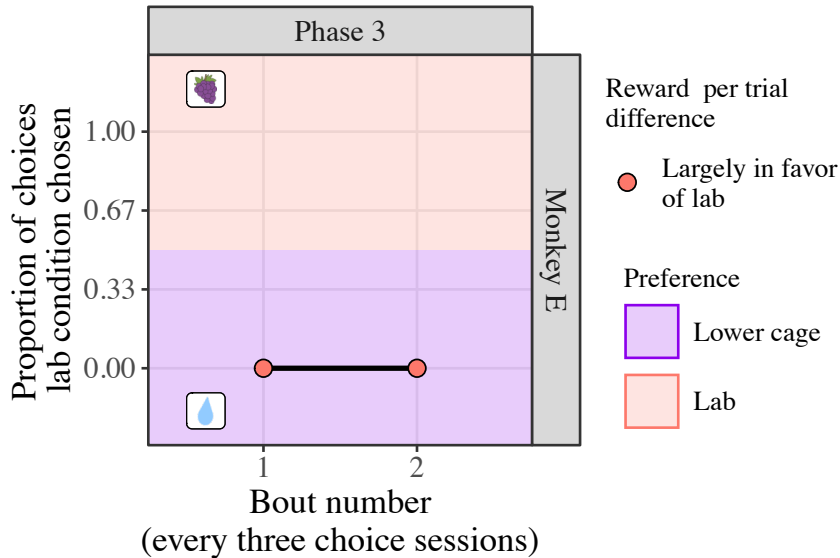


Figure 8.1: Results of the Choice-based Severity Assessment with monkey E. Proportions were calculated using three consecutive choice sessions (represented by point size) as we assessed preference and adjusted the reward per trial difference accordingly at this time. The reward per trial difference is indicated by color. One bout of testing occurred over nine days due to one reference session per condition occurring prior to for each choice session.

8.1.3.2 Model description and results of the condition stimuli training for the Choice-based Severity Assessment

We trained the monkeys to associate a visual stimulus to either receiving a 2 ml bolus from the cage condition cognitive testing system (cage condition) or once they were seated in the non-human primate chair (lab condition) prior to the Choice-based Severity Assessment (the ‘Choice’ and ‘Initial bolus’ panels of the reference sessions in Figure 2.2 of Chapter 2 depict the basic training steps). During training, each condition stimulus was paired with a timeout stimulus multiple times over a session for each monkey. Therefore, we fit three GLMMs (one per monkey) with the response variable as the number of correct responses (i.e., touching the condition stimulus) in relation to the total number of trials tested of each trial type for each session. Trial type was our main fixed effect of interest. We added session number as an additional fixed effect and session as a random effect, with all possible random slopes to allow the slopes to vary across sessions (Barr et al., 2013; Schielzeth & Forstmeier, 2009).

There was substantial evidence that all monkeys were able to differentiate the condition stimuli from the unrewarded stimulus, where monkey D also performed substantially better for upper cage than lab condition stimulus trials (Figure 8.2; Table 8.1). Generally, performance on cage condition trials was higher on average than lab condition stimulus trials (monkey H, upper cage: 0.99 % \pm 0.02 %; monkey H, lab: 0.72 % \pm 0.31 %; monkey D, cage: 1.00 % \pm 0.01 %; monkey D, lab: 0.78 % \pm 0.08 %; monkey E, lower cage: 0.96 % \pm 0.10 %; monkey E, lab: 0.82 % \pm 0.19 %). Overall, these results validate that the monkeys understood the link between the visual stimuli and their associated conditions during the Choice-based Severity Assessment.

Additionally, these findings suggest that the monkeys preferred to receive a reward from the cage condition cognitive testing system than in the non-human primate chair associated with the lab condition. Such preferences may be driven by the differences in movement constraint between the two training trial types. Given that movement constraint is the first of other expected costs to lab condition (i.e., sociality, transport; see Figure 2.1 and Table 2.1 in Chapter 2), preferences for the cage condition during the Choice-based Severity Assessment would be expected.

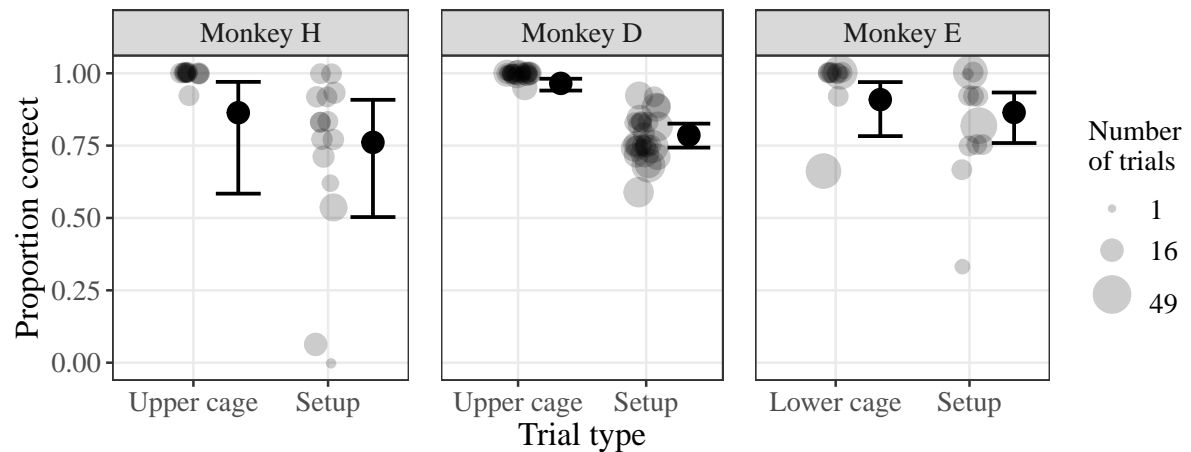


Figure 8.2: Results of the condition stimuli training for the Choice-based Severity Assessment. The light grey points indicate the proportion of trials each condition stimulus was chosen (i.e., correct response) over the total number of trials conducted for that stimulus type for each monkey ($N = 3$) and session they participated in (range: 2 to 44 trials). The black points indicate the model probability estimates for each trial type. Whiskers represent the 95 % credible intervals.

Table 8.1: Model results of the condition stimuli training for the Choice-based Severity Assessment. Binomial generalized linear mixed models (one per monkey) tested whether the correct stimulus was chosen (i.e., condition stimulus).

Monkey	Variable	Estimate	SD	Lower CI	Upper CI	Pr
H	Intercept	1.90	0.37	1.15	2.64	1.00
	Trial type (cage) ^a	0.52	0.44	-0.38	1.34	0.88
	Trial number	-0.53	0.40	-1.29	0.30	0.91
D	Intercept	1.31	0.13	1.06	1.56	1.00
	Trial type (cage) ^a	2.05	0.30	1.47	2.64	1.00
	Trial number	-0.01	0.15	-0.32	0.28	0.53
E	Intercept	1.90	0.37	1.15	2.64	1.00
	Trial type (cage) ^a	0.52	0.44	-0.38	1.34	0.88
	Trial number	-0.53	0.40	-1.29	0.30	0.91

Estimate: slope of the predictor. SD: standard deviation of the estimate. CI: 95 % credible interval. Pr: proportion of the posterior samples that fall on the same side of 0 as the mean.

^aLab was the reference level for trial type.

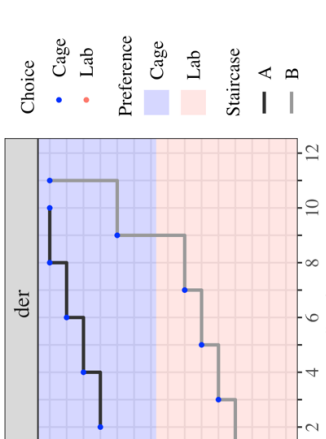
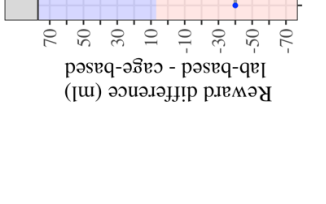
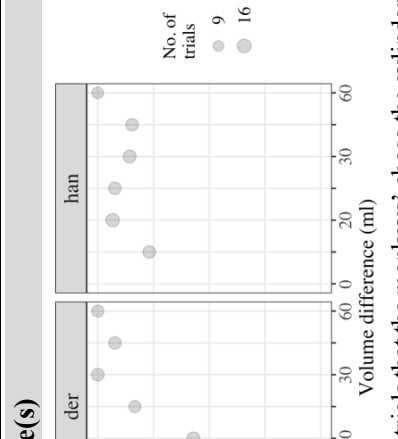
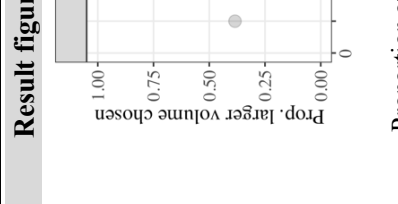
8.1.4 Supplementary video

A video describing the Choice-based Severity Scale and the procedure of the Choice-based Severity Assessment will be included upon publication of the manuscript.

8.1.5 Supplementary experiments

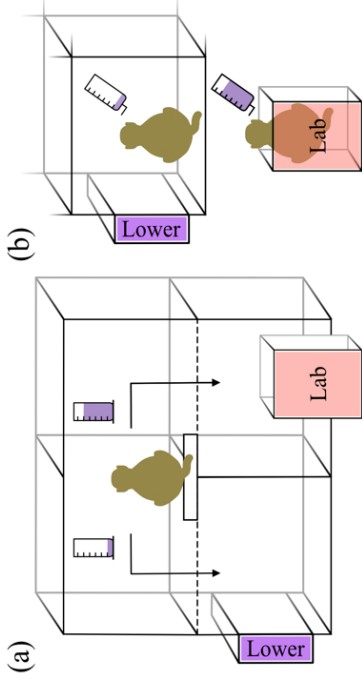
Table 8.2: Compilation of the pilot experiments that informed the design of the Choice-based Severity Assessment (see next page).

Table 8.2: Compilation of the pilot experiments that informed the design of the Choice-based Severity Assessment.

Choice-based Severity Assessment: Vertically positioned settings	Result figure(s)	Main conclusions
<p data-bbox="327 185 391 481">Purpose: First testing structure</p> <div data-bbox="327 481 654 918">  </div> <p data-bbox="662 481 805 918">(a) Positioning of the conditions and their associated bolus as indicated by the cylinders. The vertical sliding panels were pulled simultaneously so the monkey could choose a condition by entering one of the quadrants. (b) Once the monkey made a choice, it received the corresponding bolus of the chosen condition.</p>	<div data-bbox="327 929 654 1131">  </div> <p data-bbox="662 929 805 1131">Plot of monkey's choices across sessions. The bolus to the monkey's choice using two interleaved staircases that alternated each session.</p>	<ul data-bbox="327 1142 805 2047" style="list-style-type: none"> • The monkey only chose the cage condition. • Unclear if this preference was due to the position or the option itself. • The monkey also might not have paid attention to or been able to differentiate the volumes of fluid.
<p data-bbox="813 185 877 481">Purpose: To train the monkeys to pay attention to the fluid reward cylinders and to determine their volume differentiation skills.</p> <div data-bbox="813 481 1212 918">  </div> <p data-bbox="1220 481 1458 918">(a) Once the monkey was centered between the top two quadrants, cylinders with different amounts of bolus were placed on either side for the monkey to choose between. The monkey indicated his choice by tapping the position of one of the cylinders. (b) Once the monkey made a choice, the horizontal sliding panel was opened to the corresponding quadrant. The monkey received the chosen bolus once he entered the open lower quadrant for both conditions.</p>	<div data-bbox="813 929 1212 1131">  </div> <p data-bbox="1220 929 1458 1131">Proportion of trials that the monkeys' chose the cylinder with the larger fluid reward volume. The pairs of monkey volumes presented were 0 vs. 60, 5 vs. 55, 10 vs. 50, 15 vs. 45, 20 vs. 40, and 25 vs. 35. 30 vs. 30 was also tested but is not presented in the graph.</p>	<ul data-bbox="813 1142 1458 2047" style="list-style-type: none"> • The monkeys were quite good at differentiating volumes (choosing the larger volume over 75% of trials) when the difference between the cylinders was over 25 ml. • This training was conducted in preparation for the next pilot choice experiment (horizontally positioned conditions).

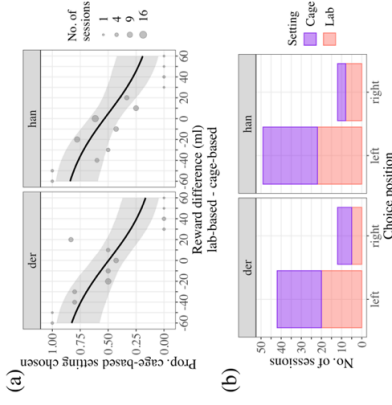
Choice-based Severity Assessment: Horizontally positioned settings

Purpose: To control for the bias due to positioning the options up and down



(a) Positioning of the conditions and their associated bolus as indicated by the cylinders. Once the cylinders were placed into position, the monkey indicated his choice by tapping the position of one of the cylinders. (b) Once the monkey made a choice, the horizontal sliding panel was opened to the corresponding quadrant. The monkey received the chosen bolus once he had condition by their position during the choice phase of the experiment.

Result figure(s)



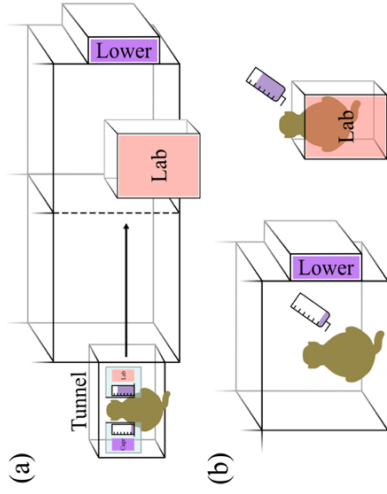
(a) Proportion of sessions that the lab condition was chosen per reward difference for each monkey. (b) Number of sessions that the monkeys chose each condition by their position during the choice phase of the experiment.

Main conclusions

- We can influence the monkeys' choice with reward (psychometric function graph).
- Monkeys exhibited a side bias (independent of the condition type).
- Potentially due to the location of the cage squeeze (needed for veterinary purposes) on the lower right compartment.

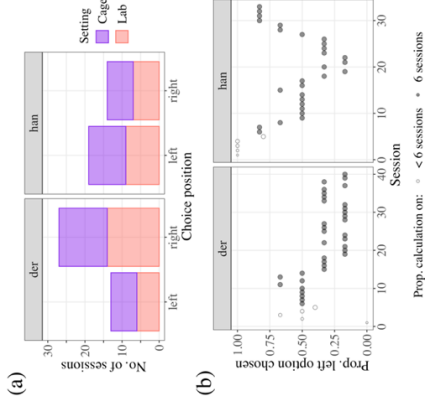
Choice-based Severity Assessment: Tunnel positioned settings

Purpose: To control for the bias due to positioning the options left and right.



(a) Positioning of the conditions and location of the tunnel where the monkey made his choice. Once the condition tags and cylinders were placed into position, the monkey indicated his choice by tapping in the condition tag. (b) The bolus for the cage condition was received in the lower right quadrant once the non-human primate chair was removed. The bolus for the lab condition was received once the monkey was seated in the chair.

Result figure(s)

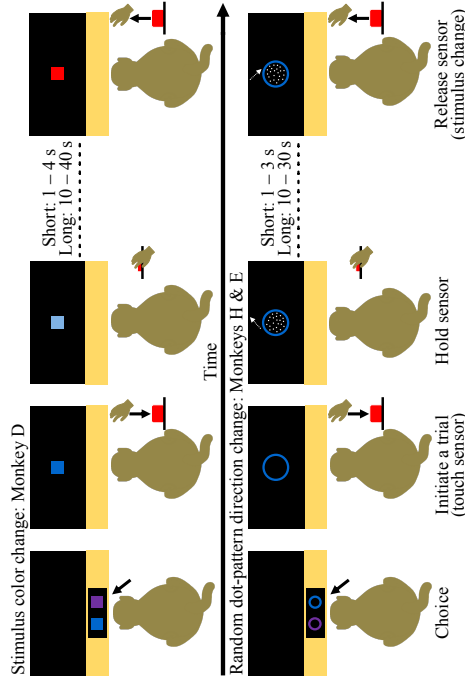


- Monkeys still prone to side biases.
- Might also not have understood that each condition stimulus and its corresponding bolus lead to different consequences.

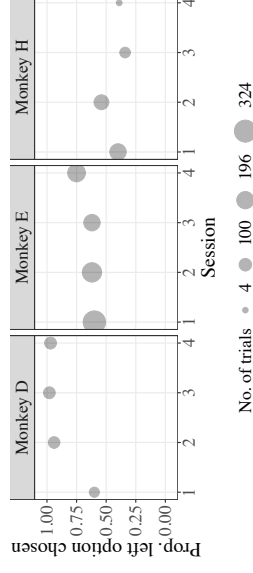
Main conclusions

Choice-based Severity Scale test: Counterbalanced options

Purpose: First testing structure of the Choice-based Severity Scale test.



Result figure(s)



Proportion of trials that the monkeys chose the option positioned on the left per session.

Main conclusions

- One individual developed a strong side bias by the second session.
- Result suggested that the monkeys could respond without an informed decision.
- Therefore, we changed the position of the options to be deterministic (see Chapter 2).

Note: For all pilot experiments where the monkeys were offered a choice between performing the basic experimental task in the cage and lab conditions, the task was started 10 minutes later to account for the time needed to transport the monkey to the lab condition. The basic experimental task ran for 2 hours in both conditions. Contact the first-author of the study for more specific information on how the pilot experiments were carried out.

8.2 Appendix B: Supplementary material for Chapter 3 (Long-delay learning in non-human primates)

8.2.1 Supplementary information

To investigate the that uniformed trial might have on the monkeys' preference, we examined their choices across completed trials during the generalized stimuli experiment. We grouped the data by each monkey and stimulus set (N = 12) using a sliding window of the choices made in last 12 completed trials. We set our preference threshold criteria to be least 80 % of the last 12 completed trials (i.e., 10 or more completed trials of the same choice). Following our preference threshold criteria (i.e., at least 80 % of the last 12 completed trials), we found that the monkeys developed a preference for the high reward stimulus by the last testing session for each stimulus set (Figure 8.3). Generally, choices for the high reward stimulus of completed trials increased once the monkeys became informed of the consequences of both stimuli (Figure 8.3). Therefore, we decided to exclude uniformed trials from these analyses of the generalized experiment.

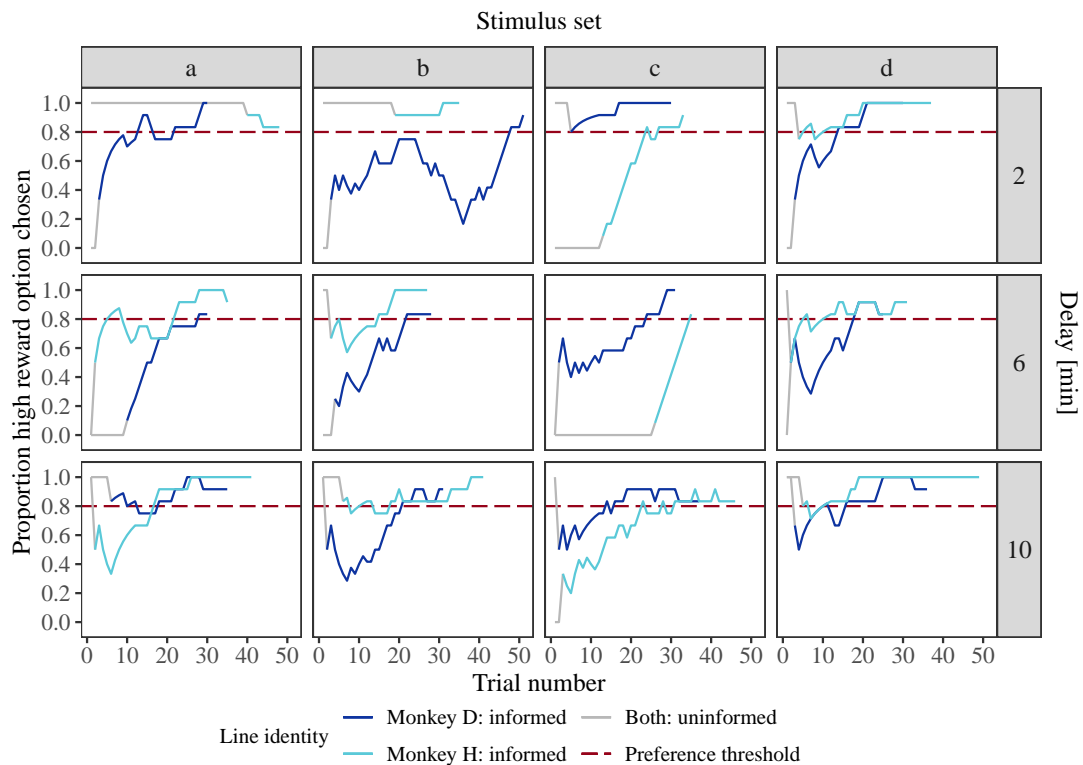


Figure 8.3: Proportion of high reward choices over the last 12 completed trials for each stimulus set of each delay in the generalized stimuli experiment. A change in line color from grey to dark blue (monkey D) or light blue (monkey H) indicates the point when the monkeys became informed of the consequences of both stimuli (i.e., trials occurring after one trial of each stimulus was completed). The dashed red line indicates the threshold that monkeys were considered to prefer the high reward stimulus.

Table 8.3: Results of the leave-one-out cross-validation for the choice behavior analyses. Models are ordered by their expected log predicted density difference (ELPD diff.) from the top-ranking model.

	ELPD	ELPD diff.	ELPD SE diff.	Cand. set	Final model
Choice model 1: Is choice modulated by increasing delay?					
1. delay*monkeyID	-445.62	0.00	0.00	✓	
2. delay + monkeyID	-445.95	-0.33	2.44	✓	✓
Choice model 2: Is choice modulated by increasing delay, despite stimuli being novel?					
1. delay + monkeyID	-630.21	0.00	0.00	✓	
2. delay*monkeyID	-630.22	-0.01	0.94	✓	✓

Models were included in the candidate (Cand. Set) for consideration when their ELPD diff. was within two times the standard error (SE) difference (negative values indicate a worse fit in comparison to the top-ranking model). The simplest model was chosen when ELPD did not differ substantially (Final model).

Table 8.4: Results of the leave-one-out cross-validation for the choice behavior analyses. Models are ordered by their expected log predicted density difference (ELPD diff.) from the top-ranking model.

	ELPD	ELPD diff.	ELPD SE diff.	Cand. set	Final model
Trial outcome model 1: Is trial outcome influenced by choice and/or increasing delay?					
1. choice*delay + choice*monkeyID	-777.07	0.00	0.00	✓	✓
2. choice*delay + choice*monkeyID + delay*monkeyID	-777.69	-0.62	0.69	✓	
3. choice*delay*monkeyID	-778.04	-0.97	0.76	✓	
4. delay + choice*monkeyID	-782.86	-5.79	1.90		
5. choice*monkeyID + delay*monkeyID	-782.99	-5.92	1.90		
6. choice*delay + monkeyID	-790.51	-13.44	3.84		
7. delay*choice + delay*monkeyID	-791.80	-14.73	3.79		
8. choice + delay + monkeyID	-796.82	-19.75	4.01		
9. choice + monkeyID*delay	-797.64	-20.57	3.94		
Trial outcome model 2: Does trial outcome influenced by choice and/or increasing delay, despite stimuli being novel?					
1. choice*monkeyID + delay	-874.99	0.00	0.00	✓	
2. choice*monkeyID + delay*monkeyID	-875.10	-0.11	1.10	✓	
3. choice*delay*monkeyID	-875.23	-0.24	1.37	✓	
4. choice + delay*monkeyID	-875.55	-0.56	1.52	✓	
5. choice*delay + choice*monkeyID + delay*monkeyID	-875.90	-0.91	1.32	✓	
6. choice*delay + choice*monkeyID	-876.72	-1.73	0.70	✓	
7. choice*delay + monkeyID	-876.80	-1.81	1.28	✓	
8. choice + delay + monkeyID	-876.89	-1.90	1.15	✓	✓
9. choice*delay + delay*monkeyID	-877.07	-2.08	1.67	✓	

Models were included in the candidate (Cand. Set) for consideration when their ELPD diff. was within two times the standard error (SE) difference (negative values indicate a worse fit in comparison to the top-ranking model). The simplest model was chosen when ELPD did not differ substantially (Final model).

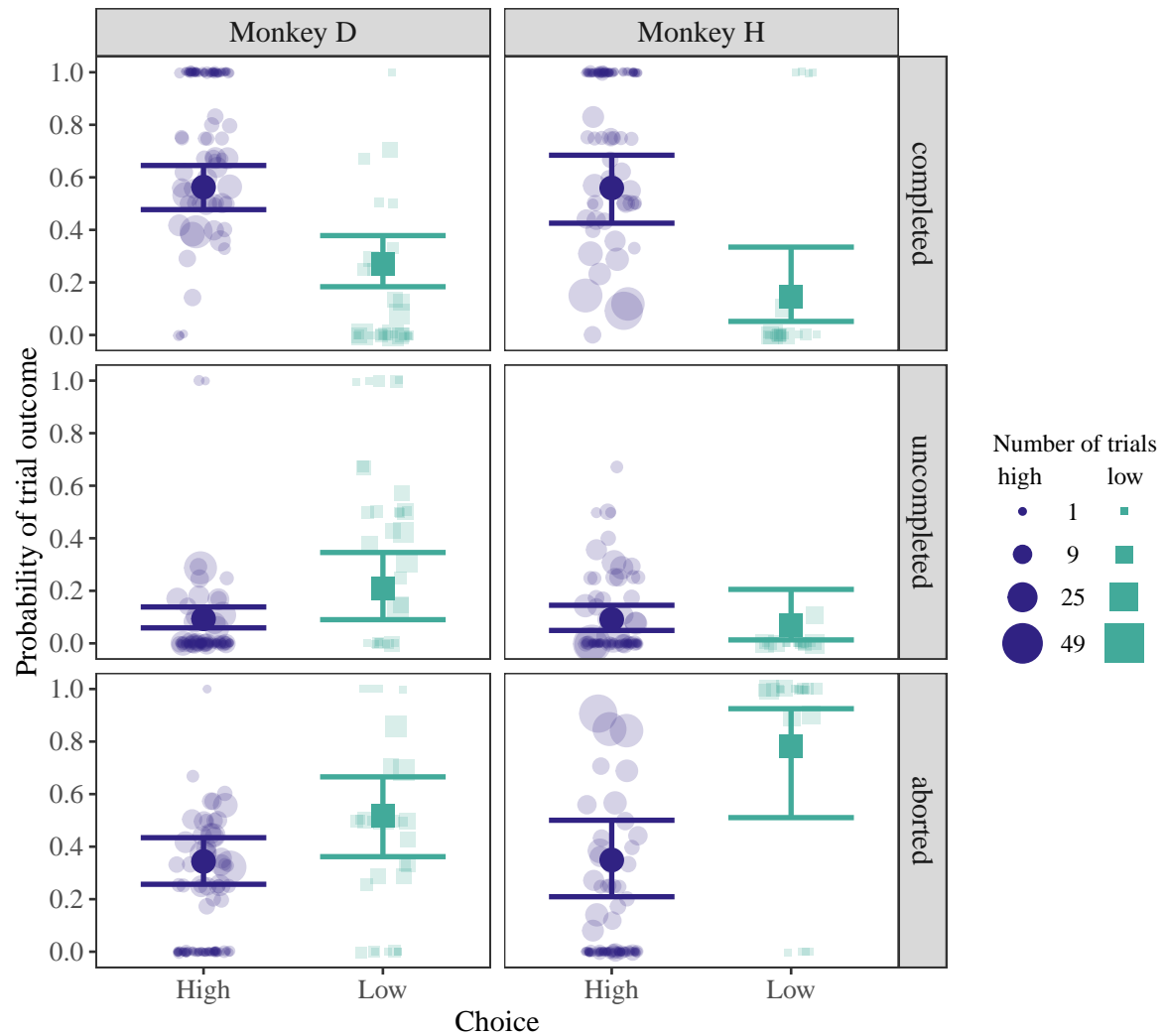


Figure 8.4: Effect of choice and animal identity on trial outcome behavior during the fixed stimuli experiment. Probability of each trial outcome (completed, uncompleted, aborted) by choice (low reward, high reward) and monkey (D, H) Point size reflects the total number of trials conducted by each monkey, delay, session, and choice (range: 1 to 43 trials; see legend). The position of each point represents the proportion of trials that were completed, uncompleted, or aborted. Dark colored points indicate the model probability estimates. Whiskers represent the 95 % credible intervals. Model probability estimates and credible intervals calculated from a model run with all other variables are at their mean (factors dummy coded).

8.1 Appendix C: Supplementary material for Chapter 3 (Dot-probe attention bias task)

8.1.1 Stimuli preparation and additional analytics

8.1.1.1 Training stimuli

Training stimuli consisted of unknown conspecific infants with or without adult females (social image, $N = 59$), and scrambled facial images of adult males (scrambled image, $N = 7$). We processed all images using Adobe Photoshop CS3 Extended Version 10.0. Social images (size 700 by 700-pixels) with the actor centered in the image and formatted on a grey square (size 800 by 800-pixels; RGB: 191, 191, 191). Male facial scrambled images were produced by first cropping the head and resizing it so that the largest dimension of the head (i.e., length or width of head) was 700-pixels. Secondly, we “closed” the eyes of each image to minimize potential aggressive content as macaques perceive eye contact as threatening (van Hooff, 1967) and eyes are salient facial substructures for these species (Guo et al., 2003). This step was performed using the stamp tool and selecting appropriately colored areas of the face to stamp over the pupils to mimic an eyelid. Then, we centered each image on an 800 by 800-pixels grey background (RGB: 191, 191, 191). We scrambled the training images using the “Scramble filter” (60 by 60-pixel squares), an open-source plug-in created for Adobe Photoshop by [Telegraphics Inc.](http://www.telegraphics.com.au/sw/) (<http://www.telegraphics.com.au/sw/>). Finally, each training image was mirrored to form a stimulus pairs. We labeled pairs of mirrored images so that the actors’ facial and/or body position would be directed towards the center of the touchscreen when presented side-by-side.

8.1.1.2 Test stimuli

We compiled test stimuli from pictures of six unknown conspecific adult males for each of whom a picture with a with neutral expression and an aggressive expression were available ($N = 6$). These images were morphed within their respective facial expression to expand the number of images in our test stimuli set to $N = 18$ (i.e., additional 12 morphs with neutral and aggressive facial expressions). We first processed the novel stimuli as described for the facial training images above (i.e., cropping the head, formatting on grey background) and modifying the eyes of only the neutral images. Then, we adjusted the luminance and color tone of each identity matched pair of neutral and aggressive stimuli to be similar. Using FantaMorph

software (version 5.4.8), we morphed by two image JPEGs of the same facial expression by adding dots to key structures of the head (e.g., eyes, nose, mouth) on one of the images and dragging the dot that appeared simultaneously on the other image to the corresponding facial feature. The resulting morphed image represented 50 % of each image included. Morphed images were processed like the original testing stimuli and original facial images in the training image set (i.e., cropping the head, formatting on grey background). After processing, we scrambled each image using a custom script written in MATLAB (version 9.0.0.341360) that randomly scrambled 10 by 10-pixel squares within the shape of the head present in the image. We adjusted squares (e.g., relocated) where necessary to ensure the image did not retain any facial features of the original image. Finally, we mirrored all testing images and matched them with their respective counterpart (e.g., individual with neutral expression and closed eyes matched with that individual with open eyes and an aggressive expression). Lastly, we labeled pairs of images so that actors' eye and/or body position in the images during the dot-probe task would be directed the towards the center of the touchscreen.

8.1.2 Supplementary analysis of low-level stimulus features

Even though the whole face images of aggressive and neutral expressions used in our experiment did not differ in low-level stimulus features (see “Stimuli preparation” section in main manuscript for analysis), they did differ in shape (i.e., outline of head). As these images were formatted on a grey square background, these differences in shape could result in more or less of the background being visible. Such differences in background visibility could also change aspects of the formatted stimuli (cropped head on grey square) given that they were presented on a black screen during the dot-probe task. Therefore, we further investigated characteristics of the formatted stimuli as our findings indicated that scrambled stimulus pairs also elicited a similar pattern of attention bias as whole face stimulus pairs.

All images were analyzed in MATLAB (version 9.5.0.1298439; Bradley et al., 2007). Each image was read into MATLAB using the “imread” function. Color values were calculated individually by averaging the respective values across all image pixels. Luminance was calculated as the mean RGB value for each pixel, averaged across all pixels (Bradley et al., 2007). The standard deviation of the mean RGB values was computed for each column of pixels; contrast was calculated by taking the standard deviation of this first calculation (Bradley et al., 2007). For each low-level feature, we ran an LMM using the R package ‘lme4’ (version 1.1-26) with the interaction of facial expression (aggressive or neutral) and stimulus type

(whole face or scrambled) as the predictors and actor identity as a random effect (with all possible random slopes). We compared each model to its null counterpart, lacking any predictors, by using likelihood ratio tests (LRTs) via the ‘anova’ function with the argument ‘test’ set to “Chisq.” Then we used the drop1 function in a step-wise manner to determine if interactions should be retained and to obtain p-values for remaining predictors once the final models were deduced.

All low-level feature models differed significantly from their null counterparts (red: $\chi^2 = 7.78$, $df = 3$, $p = 0.051$; green: $\chi^2 = 8.78$, $df = 3$, $p = 0.032$; blue: $\chi^2 = 8.66$, $df = 3$, $p = 0.034$; luminance: $\chi^2 = 8.51$, $df = 3$, $p = 0.037$; contrast: $\chi^2 = 37.91$, $df = 3$, $p < 0.001$). Specifically, aggressive stimuli were higher in color (red: $\chi^2 = 7.66$, $df = 1$, $p = 0.006$; green: $\chi^2 = 8.78$, $df = 1$, $p = 0.003$; blue: $\chi^2 = 8.66$, $df = 1$, $p = 0.003$), luminance ($\chi^2 = 8.5$, $df = 1$, $p = 0.004$), and contrast ($\chi^2 = 22.32$, $df = 1$, $p = <0.001$) than neutral stimuli (Table 8.4; Figure 8.6). Additionally, contrast was higher for whole face stimuli than scrambled stimuli ($\chi^2 = 21.97$, $df = 1$, $p = <0.001$; Table 8.4; Figure 8.6).

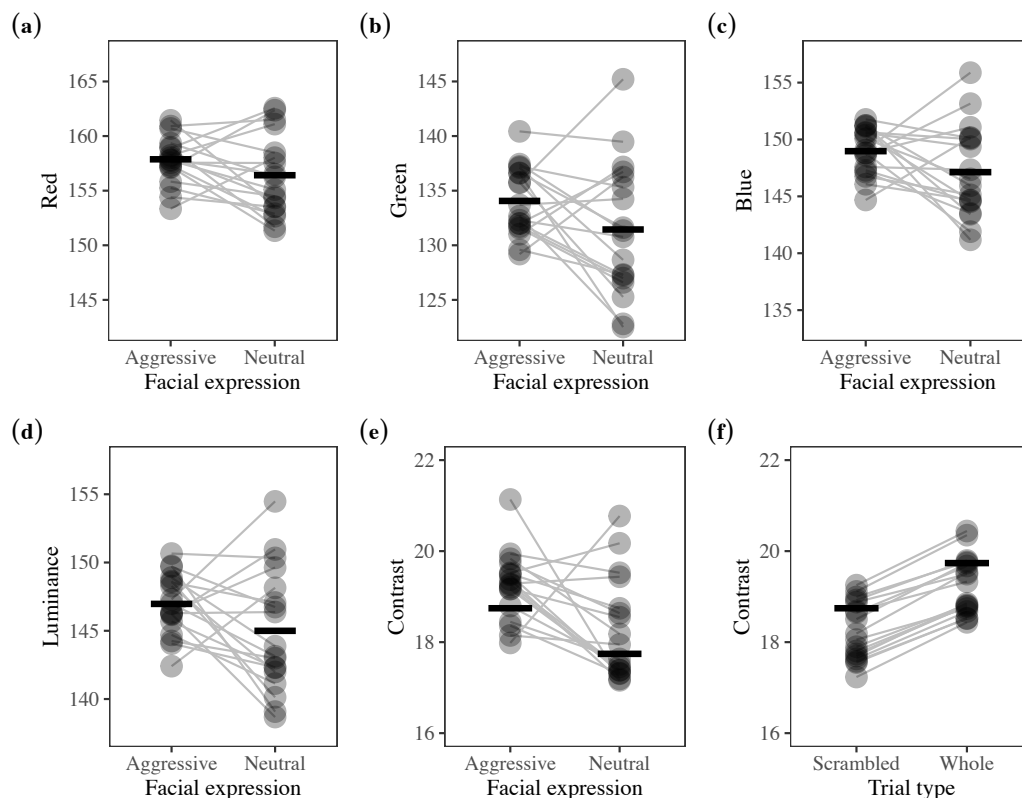


Figure. 8.6: Each point represents the average value of the corresponding low-level feature for each actor identity for scrambled and whole face formatted stimuli for plots (a), (b), (c), (d), and (e), and for aggressive and neutral formatted stimuli for plot (f)

Table 8.4: Model results for models analyzing low-level features between stimulus pairs.

Model	Estimate	SE	<i>t</i>	df	χ^2	<i>p</i> -value
Red						
Intercept	157.87	0.64	245.47			
Trial type (whole face) ^a	-0.17	0.51	-0.34	1	0.12	0.732
Facial expression (neutral) ^b	-1.46	0.51	-2.87	1	7.66	0.006
Green						
Intercept	134.06	1.03	129.85			
Trial type (whole face) ^a	-0.05	0.85	-0.06	1	0.00	0.951
Facial expression (neutral) ^b	-2.62	0.85	-3.09	1	8.78	0.003
Blue						
Intercept	148.97	0.67	223.97			
Trial type (whole face) ^a	0.04	0.60	0.06	1	0.00	0.953
Facial expression (neutral) ^b	-1.85	0.60	-3.06	1	8.66	0.003
Luminance						
Intercept	146.97	0.74	198.01			
Trial type (whole face) ^a	-0.06	0.65	-0.10	1	0.01	0.922
Facial expression (neutral) ^b	-1.97	0.65	-3.03	1	8.50	0.004
Contrast						
Intercept	18.74	0.20	95.31			
Trial type (whole face) ^a	0.99	0.19	5.21	1	21.97	<0.001
Facial expression (neutral) ^b	-1.00	0.19	-5.26	1	22.32	<0.001

^aTrial type was dummy coded with the whole face stimuli being the reference category. ^bFacial expression was dummy coded with neutral expressions being the reference category. SE: standard error; df: degrees of freedom.

Because the differences between formatted stimuli were driven by the amount of visible grey square/background, we checked what the impact of those differences on reaction time to scrambled stimulus pair trials was. For each stimulus pair, we calculated the difference in each low-level feature between the aggressive stimulus and the neutral stimulus depending on congruency (congruent: aggressive - neutral; incongruent: neutral - aggressive). We used an information theoretic approach to compare LMMs with reaction time as the response and the predictor being one of the low-level feature differences to a null model, lacking predictors for

each stimuli duration (similar to hand-preference analyses). Each model included for the interaction of congruency and dot-probe position, time tested, rank, and trial number as control variables and animal identity as a random effect (with all possible random slopes).

We found that the null model was the best model for both stimuli durations, with Akaike weights over 0.50 (100 ms: 0.57; 1,000 ms: 0.53) and the other models differing from these models by more than 2 (Table 8.5; Burnham et al., 2002). Therefore, none of the differences in low-level features explained reaction time to dot-probes following scrambled stimulus pairs. These findings led us to conclude that the attention bias effects for scrambled stimuli are likely driven another factor, such as the outline of the head being discernible.

Table 8.5: Model comparison checking for the influence of low-level features on reaction time to the dot-probe for scrambled stimulus pairs.

	AICc	Model likelihood	Δ AICc	AIC weight	Cumulative weight	Evidence ratio
100 ms models						
Null	-3261.68	1.00	0.00	0.57	0.57	1.00
Contrast	-3258.07	0.16	3.61	0.09	0.67	6.09
Red	-3257.49	0.12	4.20	0.07	0.74	8.17
Luminance	-3257.40	0.12	4.28	0.07	0.80	8.54
Green	-3257.38	0.12	4.30	0.07	0.87	8.54
Blue	-3257.36	0.12	4.32	0.07	0.94	8.67
Volume	-3257.27	0.11	4.41	0.06	1.00	9.08
1,000 ms models						
Null	-2931.16	1.00	0.00	0.53	0.53	1.00
Contrast	-2928.50	0.26	2.67	0.14	0.67	3.79
Volume	-2927.32	0.15	3.84	0.08	0.75	6.79
Green	-2927.03	0.13	4.13	0.07	0.81	7.91
Luminance	-2926.92	0.12	4.25	0.06	0.88	8.41
Blue	-2926.90	0.12	4.26	0.06	0.94	8.41
Red	-2926.77	0.11	4.39	0.06	1.00	8.98

AICc: corrected Akaike Information Criterion; Δ AICc: difference in AICc value from that of best model; Model likelihood: relative likelihood of the model given the data; AIC weight: Akaike model weight, probability that the model is the best of the set; Evidence ratio: weight of the best model divided by the weight of the given model.

8.1.3 Dot-probe task design

Both the white circular start button and dot-probe were 5.4 cm in diameter, with a touch-response area of 7.9 cm in diameter. The start button appeared 7.6 cm below the center of the touchscreen, within a black box centered on a blue bar. Stimulus pairs appeared 3.8 cm above and 6.4 cm on either side of the touchscreen center (as measured from the center of each stimulus), so that the edges of the stimuli were 3.2 cm apart. All stimuli appeared 7.2 cm by 7.2 cm in size and centered on 11.0 cm by 11.0 cm grey squares within the dot-probe task. Monkeys were able to touch the stimuli while they appeared on the touchscreen without disrupting the trial. Dot-probes would appear in the same positions as the stimulus pairs, centered 3.8 cm above and 6.4 cm on either side of the touchscreen center. If the dot-probe was touched correctly, monkeys were rewarded with a 0.25 ml drop of diluted grape juice (50 % water, 50 % grape juice). A time penalty of 750 ms was added if the monkeys touched the dot incorrectly (i.e., background touched) or did not touch the dot-probe throughout the 10 s it was present. Sound feedback (i.e., secondary reinforcer) of different tones occurred when monkeys initiated a trial (beep), touched the dot-probe incorrectly (buzz), or touched the dot-probe correctly (ding).

Each test session always began with a warm-up block in which the monkeys had to correctly touch the dot-probe replacing one stimulus after a pair of grey filler stimuli were shown for 1,000 ms. The warm-up block was followed by 3 test blocks (54 trials each), which each consisting of 2 grey filler pairs, 1 scrambled filler pair, 6 test stimulus pairs (including 2 original and 4 morphed actors) and their respective scrambled stimulus pair counterparts (12 stimulus pairs total), and 3 social filler pairs (Table 8.6). Blocks always occurred in the same order and began with 2 trials of grey filler stimuli, followed by 1 trial of scrambled filler stimuli, all shown for 1,000 ms (Table 8.6). Each block split whole face stimulus pairs and their respective scrambled counterparts equally, showing them for 100 ms or 1,000 ms (same duration every session). Each stimulus pair was counterbalanced by side and dot-probe position (i.e., congruent, incongruent) so that each pair was shown four times over the course of the block, totaling 48 trials in each block. Trial order of whole face and scrambled stimulus pairs was pseudorandomized so that a maximum of two trials of the same trial type were shown in a row. Each block ended with three trials of social filler stimuli shown for 1,000 ms. The program ran without a time limit, ending once the last test block was complete or if terminated by the experimenter (N = 1 test session for monkey E).

Table 8.6: Dot-probe task trial structure.

Block identity	Stimulus pair type	Number of trials	Trial number	Duration (ms)
Warm-up	Grey filler	1	1	1,000
Test block 1	Grey filler	2	2-3	1,000
	Scrambled filler	1	4	1,000
	Scrambled control	24	5-52	100, 1,000
	Whole face	2	5-52	100, 1,000
	Social filler	3	53-55	1,000
Test block 2	Grey filler	2	56-58	1,000
	Scrambled filler	1	59	1,000
	Scrambled control	24	60-107	100, 1,000
	Whole face	24	60-107	100, 1,000
	Social filler	3	108-111	1,000
Test block 3	Grey filler	2	112-114	1,000
	Scrambled filler	1	115	1,000
	Scrambled control	24	116-164	100, 1,000
	Whole face	24	116-164	100, 1,000
	Social filler	3	165-167	1,000

Whole face and scrambled control stimulus pairs within each block consist of 6 actor identities (3 per stimuli duration) that are shown 4 times each so that they are counterbalanced by the position on the touchscreen and dot-probe congruency. These stimuli are pseudorandomized amongst each other within a block.

8.1.4 Dot-probe training task and procedure

All monkeys were naïve towards using the touchscreen system and receiving diluted juice reward (50 % grape juice, 50 % water) as a positive reward at the start of training. We first trained the monkeys to associate touching the touchscreen with receiving a juice reward. To achieve this step, we used a combination of an automatized training task (described in Berger et al., 2018) and increasing the size of the start button (located on the lower third of the touchscreen) of the dot-probe training task (similar to the final experiment; see Fig. 2). The dot-probe training task included the training stimuli set, sound feedback, and flexibility to change experimental parameters such as dot-probe size, stimuli duration, and reward drop size. Both training tasks

began by providing the monkeys with a large stimulus (automatized training task: square; dot-probe training task: circle) which covered approximately half of the touchscreen. Monkeys were rewarded with diluted juice (drops ranging between 0.25 ml and 0.5 ml) for touching the large stimulus. Monkeys were able to learn the association between touching the screen and receiving juice reward in 3 to 21 sessions (mean \pm standard deviation: 10.50 ± 6.50 sessions per animal) and were trained alone (21.06 ± 9.75 min) or in the presence of group members (460.56 ± 458.61 min).

Next, we trained the monkeys to touch the dot-probe following the training stimuli using the dot-probe training task. Sound feedback served as secondary reinforcement to provide the monkeys feedback on their performance. Additionally, reward could be provided for initiating a trial and for correctly touching the dot-probe. Initially, we set the drop size for touching the start button to be lower than the reward drop size for the dot-probe (e.g., 0.2 ml/drop and 0.5 ml/drop respectively), and eventually removed the first reward entirely once the monkeys knew how to initiate a trial and touch the subsequent dot-probe. Training stimuli were presented for a randomized duration selected within a set minimum and maximum value (e.g., between 100 ms and 2000 ms). Once monkeys were able to successfully complete a trial (initiate trial plus touch the dot-probe to gain a juice reward), the maximum stimuli duration was slowly increased to be at least 2000 ms in duration. Monkeys were trained until they were able to successfully complete 80 % of trials within a training session usually lasting 20 min. Monkeys were able to learn the dot-probe training task in 9 to 29 sessions (17.00 ± 7.96 sessions per animal) and were trained alone (on average 37.62 ± 18.23 min) or in the presence of group members for (on average 53.14 ± 22.93 min).

8.1.5 Influence of hand preference on reaction time

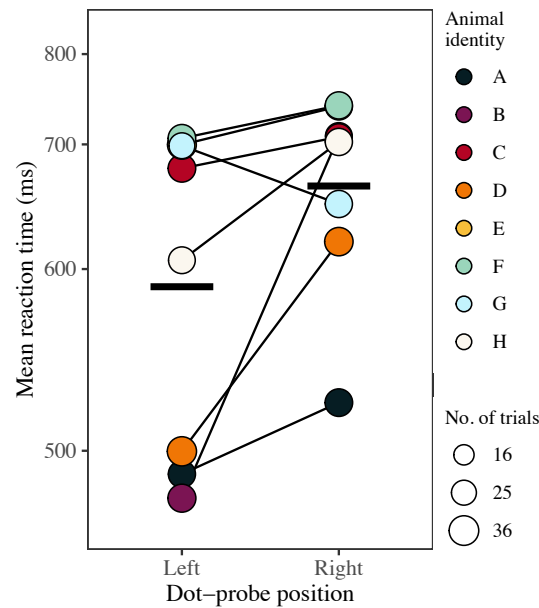


Figure 8.7: Effect of dot-probe position (control variable) on responses for 100 ms trials during the baseline test session. Each point represents the mean reaction time per monkey to dot-probes appearing on the left and right, connected by a thin black line. Point size indicates the number of trials per condition, ranging from 27 to 35 trials. Model estimates are indicated by the thick horizontal lines. Raw data are plotted per monkey in “Supplementary Figures of Raw Data.”

Table 8.7: Model comparison checking for the influence of hand preference. Full model of the baseline test session including dot-probe position (left or right) was compared to the same model including side preferred hand (ipsilateral or contralateral) instead for each stimuli duration.

	AICc	Δ AICc	Model likelihood	AIC weight	Cumulative weight	Evidence ratio
100 ms stimuli duration						
Dot-probe position	-8154.95	0.00	1.00	0.95	0.95	1.00
Position preferred hand	-8149.15	5.80	0.05	0.05	1.00	18.20
1,000 ms stimuli duration						
Position preferred hand	-7466.01	0.00	1.00	0.99	0.99	1.00
Dot-probe position	-7457.20	8.81	0.01	0.01	1.00	81.81

AICc: corrected Akaike Information Criterion; Δ AICc: difference in AICc value from that of best model; Model likelihood: relative likelihood of the model given the data; AIC weight: Akaike model weight, probability that the model is the best of the set; Evidence ratio: weight of the best model divided by the weight of the given model.

8.1.6 Other supplementary tables

Table 8.8: Dot-probe studies carried out in NHPs using affective stimulus pairs.

Article	Study species	Subjects	Stimuli	Stimulus pairs	Stimulus duration	Response window	Data included	Analysis response	Treat.	Result
King et al., (2012)	Rhesus macaques	6M	Faces	A-N	1,000 ms	Until touched, up to 60 s	RT within mean ± 2 SD ^d	AB score	T	Vigilance for threat, no change after treatment
Kret et al., (2016)	Bonobos	4F	Faces, whole body, other animals	A-N	300 ms	Until touched	Trials without interruption ^a , RT under mean + 3 SD ^b	RT	None	Vigilance for affect, particularly socio-positive
(Kret et al., (2018)	Chimpanzees	6F, 2M	Whole body, scrambled	A-A, A-N, S-S	33 ms, 300 ms	Until touched	RT under 2,500 ms	RT	None	No attention bias
Lacreuse et al., (2013)	Rhesus macaques	6M	Faces	A-N	1,000 ms	Until touched, up to 60 s	RT within 100 ms and 1,000 ms, within mean ± 2 SD ^e	AB score	None	Vigilance for threat
(Morin et al., (2019)	Rhesus macaques	12F, 13M	Faces, nonsocial	A-N	500 ms	Until touched, up to 5 s	RT within 200 ms and 1,500 ms	RT and AB score	14 subjects MALT	MALT subjects slower to respond than controls; no attention bias
Parr et al., (2013)	Rhesus macaques	2F, 4M	Faces, scrambled	A-S, N-S, DG-AG	500 ms	Until touched	Sessions with 80 % completed trials, RT under 1,500 ms	AB score	OT	Vigilance for threat, change to avoidance after treatment
Schino et al., (2020)	Tufted capuchins	5F, 5M	Grooming, no grooming	A-N	250 ms, 1,000 ms	Until touched	RT within 100 ms and 1,000 ms	RT	None	Males vigilant to grooming for 1,000 ms trials
Wilson & Tomonaga, (2018)	Chimpanzees	6F, 2M	Faces, scrambled	A-N, A-S	150 ms	Until touched	RT within 150 ms and 5,000 ms, under mean + 2 SD ^e	RT	None	No attention bias

Treat.: treatment; AB: attention bias; F: females; M: males; OT: oxytocin; RT: reaction time; T: testosterone; MALT: maternal maltreatment; SD: standard deviation; A-A: affective and affective; A-N: affective and neutral; A-S: affective and scrambled; S-S: scrambled and scrambled; N-S: neutral and scrambled; DG-AG: direct gaze and averted gaze. ^aInterruptions included nose wipes and social interference as determined by video analysis. ^bCalculation conducted per individual after first outliers removed. ^cCalculation conducted per individual, condition, and session after first outliers removed. ^dCalculation conducted on all data after first outliers removed. ^eCalculation conducted using generalized linear mixed models; attention bias analyses conducted using repeated measures ANOVAs.

Table 8.8: Descriptive statistics for each animal. Animal identity, age at the beginning of the study, the animal's hand preference, mean reaction time (ms) \pm standard deviation (first row) per test session and stimuli duration, and number of trials per test session and stimuli duration (second row, in parentheses). Monkey B did not participate in any trials at test session A + 1d (dashes).

ID	Age	Rank	Hand	pref.	100 ms					1,000 ms				
					Baseline	A + 1d	A + 3d	A + 7d	A + 14d	Baseline	A + 1d	A + 3d	A + 7d	A + 14d
A ^a	17.35	1	left		510 \pm 41 (61)	670 \pm 122 (52)	597 \pm 61 (69)	501 \pm 55 (66)	502 \pm 67 (65)	653 \pm 259 (62)	756 \pm 91 (60)	663 \pm 59 (63)	606 \pm 76 (64)	594 \pm 89 (60)
B ^b	19.13	1	left		620 \pm 213 (61)	–	941 \pm 384 (67)	758 \pm 237 (61)	785 \pm 259 (63)	682 \pm 237 (61)	–	878 \pm 321 (65)	702 \pm 255 (60)	728 \pm 191 (59)
C	7.10	3	right		732 \pm 190 (59)	802 \pm 179 (48)	731 \pm 218 (52)	626 \pm 182 (51)	667 \pm 125 (56)	686 \pm 243 (65)	739 \pm 206 (56)	659 \pm 105 (61)	564 \pm 128 (61)	622 \pm 126 (57)
D	6.59	4	left		566 \pm 92 (68)	545 \pm 83 (65)	522 \pm 84 (65)	479 \pm 57 (67)	508 \pm 86 (68)	560 \pm 163 (52)	577 \pm 197 (53)	656 \pm 256 (62)	554 \pm 128 (62)	607 \pm 143 (55)
E ^c	18.50	2	left		735 \pm 111 (64)	661 \pm 155 (53)	596 \pm 105 (65)	560 \pm 111 (66)	504 \pm 58 (65)	502 \pm 84 (52)	600 \pm 266 (32)	587 \pm 274 (33)	499 \pm 88 (34)	551 \pm 168 (34)
F	7.65	4	left		760 \pm 174 (61)	891 \pm 367 (48)	687 \pm 192 (60)	776 \pm 254 (55)	775 \pm 187 (62)	787 \pm 245 (52)	861 \pm 295 (44)	706 \pm 223 (59)	766 \pm 213 (48)	712 \pm 275 (59)
G	7.16	5	right		740 \pm 262 (57)	653 \pm 105 (44)	636 \pm 102 (62)	593 \pm 89 (60)	653 \pm 130 (58)	705 \pm 187 (57)	691 \pm 142 (37)	668 \pm 155 (46)	673 \pm 186 (59)	714 \pm 176 (52)
H	7.15	2	left		687 \pm 190 (59)	741 \pm 401 (63)	509 \pm 54 (64)	569 \pm 93 (67)	527 \pm 80 (68)	722 \pm 190 (53)	755 \pm 187 (61)	580 \pm 116 (63)	676 \pm 186 (51)	618 \pm 159 (53)

Summary data for ages at the beginning of the study, overall mean reaction time per stimuli duration, and overall mean attention bias scores for 100 ms trials are reported in the text. ^aData reported for monkey A for the A + 1d test session was collected two days following prolonged anesthesia. ^bMonkey B refused to participate in the dot-probe task one day following prolonged anesthesia (indicated by dashes). ^cThe dot-probe test at A + 1d for monkey E was stopped early (explained in main text); we report summary data for monkey E for 1,000 ms trials although we excluded her from the analysis of these (explained in text).

Table 8.9: Model analyzing attention bias score for 100 ms trials during the baseline and A + 1d test sessions. Results of the general linear mixed model examining the effect of test session on reaction time to the dot-probe.

	Estimate	SE	<i>t</i>	df	χ^2	<i>p</i> -value
Intercept	4.90e-06	3.21e-06	1.53			
Test predictors						
Test session (A + 1d) ^a	-9.91e-06	4.78e-06	-2.07	1	3.66	0.056
Control predictors						
Time tested ^b	-7.63e-07	2.39e-06	-0.32	1	0.09	0.767
Rank ^c	-7.99e-08	2.20e-06	-0.04	1	0.00	0.971

Estimates and standard error (SE) are written in scientific notation to facilitate readability. Control predictor results are also shown. ^aTest session was dummy coded with the baseline test session being the reference category. ^bTime tested was z-transformed, original values ranged from 9 to 190 minutes after 12:00, mean \pm standard deviation: 100.71 \pm 48.14 minutes after 12:00. ^cRank was z-transformed, original values ranged from 1 to 5, mean \pm standard deviation: 3.0 \pm 1.36. Note: full-null model comparison was no longer significant when a strong responder (Monkey F) was experimentally removed (LRT: $\chi^2 = 2.52$, *df* = 1, *p* = 0.113).

8.1.7 Supplementary figures of raw data

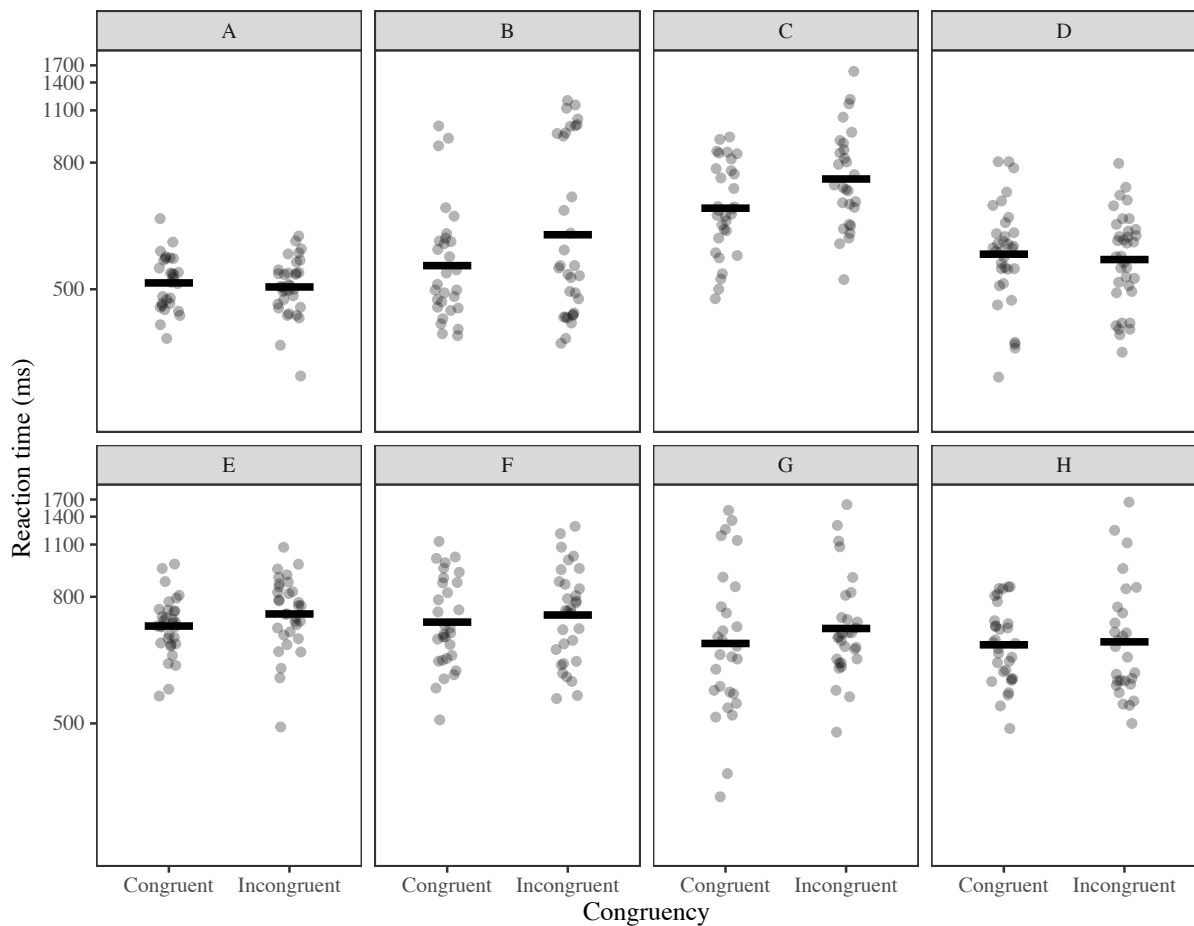


Figure 8.8: Reaction time data per monkey based on trial type for 100 ms trials during the baseline test session. Each panel corresponds to a different monkey. Each point represents one trial. The Y-axes are scaled according to the transformed data. Thick horizontal lines indicate the monkeys' mean reaction time to each condition.s

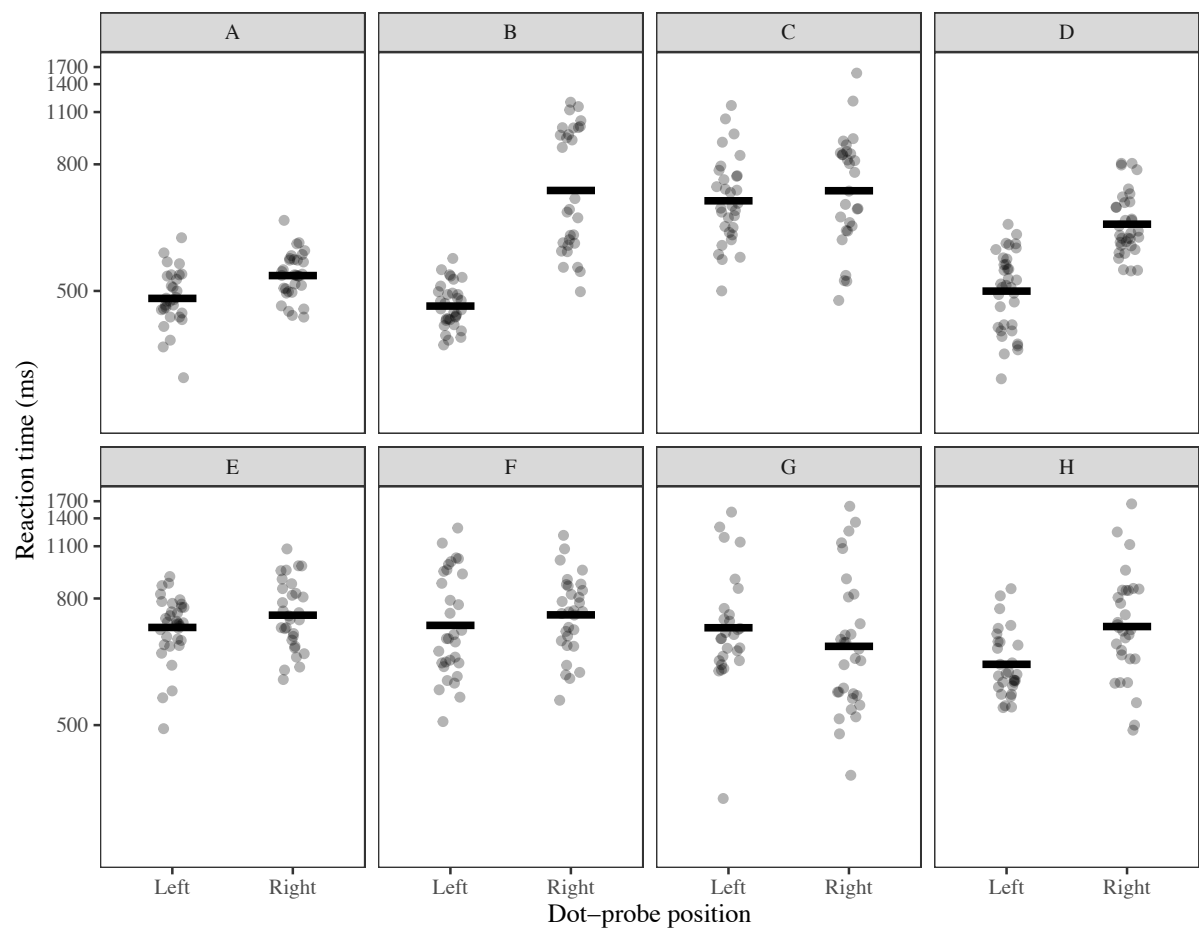


Figure 8.9: Reaction time data per monkey based on dot-probe position (control variable) for 100 ms trials during the baseline test session. Each panel corresponds to a different monkey. Each point represents one trial. The Y-axes are scaled according to the transformed data. Thick horizontal lines indicate the monkeys' mean reaction time to each condition

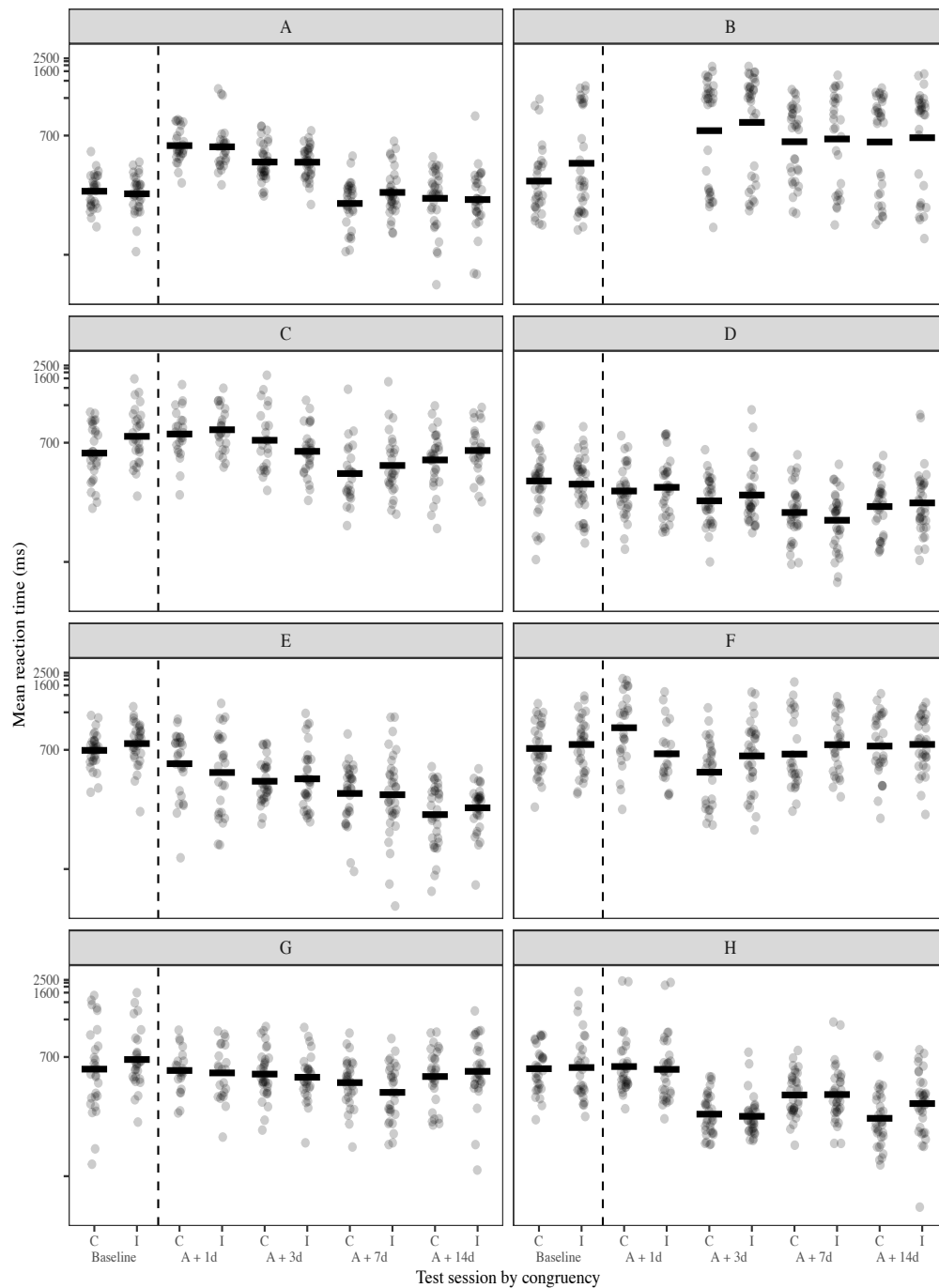


Figure 8.10: Reaction time data per monkey for the interaction of test session and congruency for 100 ms trials for each test session. Each panel corresponds to a different monkey. Monkey B did not participate in the test session occurring one day following prolonged anesthesia (A + 1d). The dashed vertical line separates baseline data from data collected in the two weeks immediately following anesthesia. Each point represents one trial. The Y-axes are scaled according to the transformed data. Thick horizontal lines indicate the monkeys' mean reaction time to each condition.

8.1.8 Supplementary references

- Berger, M., Calapai, A., Stephan, V., Niessing, M., Burchardt, L., Gail, A., & Treue, S. (2018). Standardized automated training of rhesus monkeys for neuroscience research in their housing environment. *Journal of Neurophysiology*, *119*, 796–807. <https://doi.org/10.1152/jn.00614.2017>
- Bradley, M. M., Hamby, S., Löw, A., & Lang, P. J. (2007). Brain potentials in perception: Picture complexity and emotional arousal. *Psychophysiology*, *44*, 364–373. <https://doi.org/10.1111/j.1469-8986.2007.00520.x>
- Burnham, K. P., Anderson, D. R., & Burnham, K. P. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed). Springer.
- Guo, K., Robertson, R. G., Mahmoodi, S., Tadmor, Y., & Young, M. P. (2003). How do monkeys view faces?—A study of eye movements. *Experimental Brain Research*, *150*, 363–374. <https://doi.org/10.1007/s00221-003-1429-1>
- King, H. M., Kurdziel, L. B., Meyer, J. S., & Lacreuse, A. (2012). Effects of testosterone on attention and memory for emotional stimuli in male rhesus monkeys. *Psychoneuroendocrinology*, *37*, 396–409. <https://doi.org/10.1016/j.psyneuen.2011.07.010>
- Kret, M. E., Jaasma, L., Bionda, T., & Wijnen, J. G. (2016). Bonobos (*Pan paniscus*) show an attentional bias toward conspecifics' emotions. *Proceedings of the National Academy of Sciences*, *113*, 3761–3766. <https://doi.org/10.1073/pnas.1522060113>
- Kret, M. E., Muramatsu, A., & Matsuzawa, T. (2018). Emotion processing across and within species: A comparison between humans (*Homo sapiens*) and chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, *132*, 395–409. <https://doi.org/10.1037/com0000108>
- Lacreuse, A., Schatz, K., Strazzullo, S., King, H. M., & Ready, R. (2013). Attentional biases and memory for emotional stimuli in men and male rhesus monkeys. *Animal Cognition*, *16*, 861–871. <https://doi.org/10.1007/s10071-013-0618-y>
- Morin, E. L., Howell, B. R., Meyer, J. S., & Sanchez, M. M. (2019). Effects of early maternal care on adolescent attention bias to threat in nonhuman primates. *Developmental Cognitive Neuroscience*, *38*, 100643. <https://doi.org/10.1016/j.dcn.2019.100643>
- Parr, L. A., Modi, M., Siebert, E., & Young, L. J. (2013). Intranasal oxytocin selectively attenuates rhesus monkeys' attention to negative facial expressions. *Psychoneuroendocrinology*, *38*, 1748–1756. <https://doi.org/10.1016/j.psyneuen.2013.02.011>
- Schino, G., Carducci, P., & Truppa, V. (2020). Attention to social stimuli is modulated by sex and exposure time in tufted capuchin monkeys. *Animal Behaviour*, *161*, 39–47. <https://doi.org/10.1016/j.anbehav.2019.12.019>
- van Hooff, J. A. (1967). The facial displays of the catarrhine monkeys and apes. In *Primate ethology* (pp. 7–68). Weidenfeld & Nicolson.
- Wilson, D. A., & Tomonaga, M. (2018). Exploring attentional bias towards threatening faces in chimpanzees using the dot probe task. *PLoS ONE*, *13*(11), e0207378. <https://doi.org/10.1371/journal.pone.0207378>

Declaration

I hereby declare that I have written this thesis entitled "Animal welfare from the animal's perspective: Tapping into their psychological experiences" independently and with no other aids or sources than quoted.

Lauren C. Cassidy
Göttingen, 2022

Curriculum Vitae

Name: Lauren C. Cassidy
Date of birth: 07 November 1988
Nationality: American

EDUCATION

2017 - 2022 **Deutsches Primatenzentrum, Göttingen, Germany**

PhD Program Behavior and Cognition

Cognitive Neuroscience Laboratory, Welfare and Cognition Group

Thesis title: "Animal welfare from the animal's perspective: Tapping into their psychological experiences"

Supervisors: Prof. Dr. Stefan Treue & Dr. Dana Pfefferle

2014 - 2015 **University of Roehampton, London, United Kingdom**

MRes Primate Biology, Behaviour and Conservation

Thesis: "A Comparison of Two Pair Housing Conditions Using Behavioral and Physiological Indices of Welfare in Laboratory Female Rhesus Macaques" (*Macaca mulatta*)

Supervisor: Prof. Dr. Stuart Semple

2008 - 2011 **University of California Davis, Davis, California, United States**

Bachelor of Science in Biological Sciences, emphasis Evolution and Ecology

PROFESSIONAL EXPERIENCE

2011 - 2014 **Laboratory Assistant**, Behavioral Management, California National Primate Research Center, Davis, California

2008 - 2010 **Therapeutics and Enrichment Assistant**, Primate Medicine and Environmental Enrichment, California National Primate Research Center, Davis, California

RESEARCH AND FIELD EXPERIENCE

2016 **Field assistant**, Behavioral Ecology, Georg-August-Universität, Salem, Germany

2016 **Intern**, Welfare and Cognition Group, Cognitive Neuroscience Laboratory, Deutsches Primatenzentrum, Göttingen, Germany

2015 **Intern**, Endocrinology Laboratory, Deutsches Primatenzentrum, Göttingen, Germany

2006 **Research Intern**, Behavioral Sciences Unit, Oregon National Primate Research Center, Beaverton, Oregon, United States

PUBLICATIONS

Cassidy, LC, Bethell, EJ, Brockhausen, RR, Boretius, S, Treue, S, Pfefferle, D (2021). The dot-probe attention bias task for measuring affect in animals: Design and analysis recommendations from a study with adult female long-tailed macaques (*Macaca fascicularis*). *European Surgical Research*. <https://doi.org/10.1159/000521440>.

Cassidy, LC, Leenaars, CHC, Rincon, AV, Pfefferle, D (2021). Comprehensive search filters for retrieving publications on non-human primates for literature reviews (`filterNHP`). *American Journal of Primatology*, 83, e23287. <https://doi.org/10.1002/ajp.23287>.

Cassidy, LC, Hannibal, DL, Semple, S, McCowan, B (2020). Improved behavioral indices of welfare in continuous compared to intermittent pair-housing in adult female rhesus macaques (*Macaca mulatta*). *American Journal of Primatology*, 82, e23189. <https://doi.org/10.3389/fpsyg.2019.01051>.

Hannibal, DL, **Cassidy, LC**, Vandeleest, J, Semple, S, Barnard, A, Chun, K, Winkler, S, McCowan, B (2018). Intermittent pair-housing, pair relationship qualities, and HPA activity in adult female rhesus macaques. *American Journal of Primatology*, 80, e22762. <https://doi.org/10.1002/ajp.22762>.

CONFERENCE PRESENTATIONS

Cassidy, LC, Leenaars, CHC, Rincon, AV, Pfefferle, D. Developing search filters for literature reviews: A case study with non-human primates (`filterNHP`). Poster and invited oral presentation delivered at the 113th Meeting of the German Zoological Society, online. September 2022.

Cassidy, LC, Pfefferle, D, Gail, A, Treue, S. Severity assessment using a choice-based relative ranking system. Poster presentation delivered at the 12th Primate Neurobiology Meeting, Göttingen, Germany. March 2019.

*Yurt, P, ***Cassidy, LC**, Barbosa Pereira, C, Kunczik, J, Czaplik, M, Treue, S, Gail, A, Pfefferle, D. Application of infrared thermography for monitoring implant margin condition in rhesus macaques (*Macaca mulatta*). Poster presentation delivered at the 12th Primate Neurobiology Meeting, Göttingen, Germany. March 2019.

Hannibal, DL, **Cassidy, LC**, Vandeleest, JJ, Semple, S, McCowan, B. Social partners mitigate inactivity levels in intermittently pair housed adult female rhesus macaques (*Macaca mulatta*). Poster presentation delivered at the 41st Meeting of the American Society of Primatologists, San Antonio, Texas, United States. August 2018.

Hannibal, DL, **Cassidy, LC**, Barnard, A, Vandeleest, J, Chun, K, Semple, S, McCowan, B. Intermittent versus continuous pair-housing in laboratory rhesus macaques (*Macaca mulatta*) and activation of the HPA-axis. Oral presentation delivered at 67th National Meeting of the American Association for Laboratory Animal Science, Charlotte, North Carolina, United States. October/November 2016.

Cassidy, LC, Semple, S, Hannibal DL, McCowan, B. Behavioural and physiological effects of housing type on laboratory house female rhesus macaques (*Macaca mulatta*). Poster presentation delivered at the 6th Meeting of the European Federation of Primatology, Rome, Italy. August 2015.

Hannibal, DL, **Cassidy, LC**, Day, A, Tatum, L, McCowan, B. Produce enrichment reduces alopecia in captive outdoor socially-housed rhesus macaques (*Macaca mulatta*). Poster presentation delivered at the 36th Meeting of the American Society of Primatologists, San Juan, Puerto Rico. June 2013.

* shared first authors

OTHER TALKS

Cassidy, LC, Bethell, EJ, Brockhausen, RR, Boretius, S, Treue, S, Pfefferle, D. Applying a dot-probe task in an emotional context in nonhuman primates. Oral presentation delivered at the Behavior and Cognition 5th PhD student retreat, Göttingen, Germany. November 2019.

Cassidy, LC, Pfefferle, D, Gail, A, Treue, S. Severity assessment using a choice-based relative ranking system. Oral presentation delivered at the Behavior and Cognition 4th PhD student retreat, Göttingen, Germany. October 2018.

TEACHING

"Measuring Behavior". Laboratory Animal Science (LAS) Course on Non-Human Primates (NHP): Spring 2019, Fall 2019, Spring 2020, Fall 2020, Spring 2021, Fall 2021, Spring 2022.

OTHER

Behavior and Cognition Ph.D program student representative. June 2018 - October 2020.

WEBSITES

<https://filterNHP.dpz.eu>