Philosophy of Science (2023), **90**, 1255–1264
doi:10.1017/psa.2023.10

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

CONTRIBUTED PAPER

Validating Indicators of Subjective Animal Welfare

Heather Browning

University of Southampton, Department of Philosophy, Southampton, UK and London School of Economics and Political Science, Centre for Philosophy of Natural and Social Science, London, UK
Email: drheatherbrowning@gmail.com

(Received 12 December 2022; accepted 10 January 2023; first published online 17 February 2023)

Abstract

Measurement of subjective animal welfare creates a special problem in validating the measurement indicators used. Validation is required to ensure indicators are measuring the intended target state, and not some other object. While indicators can usually be validated by looking for correlation between target and indicator under controlled manipulations, this is not possible when the target state is not directly accessible. In this article, I outline a four-step approach using the concept of robustness that can help with validating indicators of subjective animal welfare.

1. Introduction

In animal welfare science, the aim is to measure the welfare of animals under different conditions. Often, this is taken to be subjective, or hedonic welfare (e.g., Browning 2020; Duncan 2002; Mellor and Beausoleil 2015)—made up of positively and negatively valenced subjectively experienced mental states—and that is the type of welfare I am interested in here. Even where one takes a difference conception of animal welfare, measurement of subjectively experienced mental states will still be an important project for any who take them to form at least some part of welfare. However, there is a problem for measuring subjective welfare.

When measuring scientific targets, measurement can be performed either directly (e.g., taking measurements of weight or length), or indirectly using indicators (e.g., measuring temperature using a mercury thermometer). Use of indicators will typically occur because the target entity is one of three types: a composite or construct (e.g., socioeconomic status), a target that is difficult or costly to measure directly such that indicators are cheaper or easier to use (e.g., biodiversity), or a target that is simply not available to direct measurement—what I call a "hidden" target (see Browning 2020 for further discussion of these categories and their features). Subjective animal welfare is a prime example of a "hidden" scientific target.

We can't directly access the mental states that make up subjective welfare, for external or objective measurement. Instead, we rely entirely on indicator measures such as changes in behavior or physiology.

When using measurement indicators, it is important that they are valid—that is, that the indicator is measuring the intended target state. For something to function as an indicator, it must be the case that it reliably correlates/covaries with the underlying state that it is standing in for, and this requires a causal relationship with the target state. Sidestepping as much as possible the literature on the relation between correlation and causation, we can generally assume that when there is a reliable correlation between two variables A and B, it is either because A causes B, B causes A, or there is a common cause for both A and B. When we are looking for indicator measures, they will stand in one of these three relationships with the target state—they will either be a cause of the target, an effect of the target, or a mutual effect of a common cause. These three categories of indicators will have different features, both mathematically and pragmatically. Here I will focus on the first two categories, the causal and effect indicators, as the common cause type are likely to be much less common. Animal welfare science commonly uses both causal and effect indicators—often referred to as input and outcome measures, or environmental and animal-based indicators.

Bollen and Lennox (1991) differentiate between causal and effects indicators. Effect indicators are those that stand causally downstream from the target state. Changes in the indicator are a result of changes in the target. These indicators are then determined by the underlying state we want to measure. In animal welfare science there are physiological and behavioral indicators that are used to measure changes in welfare, where it is assumed that a change in the indicator reflects a change in the underlying subjective experience. For example, measurements of blood cortisol levels or approach and withdrawal behavior.

By contrast, causal indicators stand causally upstream from the target state, where changes in the indicator are a cause of changes in the target. The crucial difference here is that the indicators are *determining* the target variable rather than *determined by* it. Although both types of indicators will correlate with the target state, with effects indicators we are observing the effects of an underlying state, while with causal indicators we are observing the causes of that state. Some examples of the use of causal indicators are animal welfare assessments, which look at conditions for welfare—those things that will cause changes in the subjective states that comprise welfare (e.g., presence of adequate food and water, freedom from disease, or adequate mental stimulation). Importantly, these types of indicators will require different types of validation.

2. Validation

The validity of a test or measure refers to whether it is measuring what it purports to—whether the observed data are tracking the intended phenomenon, as opposed to some other state. Validation of indicators is thus testing to ensure that the indicators are tracking the right target state—that the values and changes in indicators are correlating with changes in the target. We need to establish that one of the types of causal relationships discussed in the preceding text holds between the indicator and the target. The process of validation will vary depending on what type of target we are discussing.

https://doi.org/10.1017/psa.2023.10 Published online by Cambridge University Press

For hidden targets such as subjective welfare, there is a particular problem for validating the indicators.

For many indicators, such as the cases of difficult measurement targets, validation can proceed through looking for reliable correlations between the indicator and direct measurement of the target. This involves first determining the causal direction (i.e., whether we have a causal or an effect indicator), which is typically established through embedding within a theoretical framework that explains the causal connections between the target and the indicators (Lindenmayer and Likens 2011; Bringmann and Eronen 2016)—or through testing to look for timing and direction of effect. The next step is to then establish a reliable correlation by measuring both the target and the indicator under a range of conditions and (preferably) interventions. If we see a reliable correlation between the target and indictor under a range of conditions, we have good reason to think that there is a valid causal connection. What is requires is correlation over a range of interventions (Markus and Borsboom 2013). Once a single indicator has been validated, we can either validate further indicators by also testing against the target, or through correlation with the known indicators.

This process is not possible for hidden targets such as subjective animal welfare. Here, the central change in the target cannot be measured for comparison and so must be validated another way. We cannot get correlational data between the target and the indicator because the target cannot be measured. All we can get is data about changes in the various indicators; there is no starting point at which we can connect an indicator to the target. Schickore and Coko (2013) point out that, in these cases, "a set of background assumptions is needed to describe how the unobservable entities bring about the experimental outcomes" (297). We are making assumptions about the causal link between the target and indicators, but the problem arises in justifying or testing these assumptions without access to the target. In the following section I will outline how robustness analysis can help resolve this problem and serve as a test of the assumptions.

3. A Four-Step Robustness Solution

As described in the preceding text, there is a validation problem for hidden targets such as subjective welfare: As we cannot access the target, we have no means of directly establishing a correlation between the target and the indicators. Instead, we must make some assumptions about the relationship between the target and indicators, and these assumptions may not be justified. As will be described, assumptions can be justified through theoretical plausibility, and tested through the collection of multiple independent lines of evidence that support the assumptions made—robustness analysis. I propose a four-step approach to validating indicators of hidden targets such as subjective welfare:

1. Make a (plausible) starting ***assumption*** relating a causal or effect indicator to the target.
2. Test for correlated variation in an indicator ***of the other type***.
3. Repeat tests for the indicator using ***different assumptions*** to give robust results.
4. Use validated indicators as starting point to test others.

https://doi.org/10.1017/psa.2023.10 Published online by Cambridge University Press

In the following text, I will detail what is involved in each of these steps, how they fit into the process, and how they will help with the problem of validating hidden indicators.

3.1. Make a (plausible) starting assumption

The first step in validating an indicator of a hidden target like welfare is to make a (plausible) starting assumption relating a causal or effect indicator to the target. An assumption of this type is necessary, as we cannot in the beginning have any knowledge about the relationships between the target and its indicators. Even in the standard case, as described earlier, we must still begin with a similar assumption. The difference in this case is that we are not then immediately going on to test this assumption but are instead using the assumption as a base to test other hypothesized target-indicator relationships. Whichever indicator we are assuming about, we can call the "set" indicator. In any particular test, we will hold this assumption fixed, using it as a basis to test other indicators (as described in step 2); but overall we will give some support to this assumption through use of different tests (as described in step 3).

One important feature of this step is that we want the starting assumption to be plausible. This means that we have some good reason to think the assumption is true, or at least justified, independently of the results of these tests. Plausibility of this type is usually achieved through embedding within an accepted theoretical framework; one that can give a description or explanation of the assumed causal relationship between the target and the indicator. If the theoretical framework is a well-accepted and well-supported one, we have good support for the plausibility of assumptions that fit within it. This is a role for existing data and accepted theory in the relevant area (Markus and Borsboom 2013).

In animal welfare science, this will primarily be sentience research, alongside evolutionary and behavioral biology. The relevant theoretical frameworks are scientific understanding of the neurophysiology of mental experience, as well as the mechanisms that underlie processing of causal indicators and expression of effect indicators (Beausoleil and Mellor 2017). If we understand the mechanisms working between welfare experience and the measured indicators, we have more reason to think they our measurements are mapping onto the right state of the world. So, if we take the vocalizations of goats, we will have more confidence that this is mapping onto welfare experience if we can understand that goats are social animals that communicate their distress to conspecifics. If we take blood cortisol measurements, we will be more confident with their reliability if we understand the hormonal cascade that creates changes in cortisol and under what conditions it is triggered. We will also have reason to think we have made the right choice of conditions, or causal indicators, from which to perform our tests. For example, understanding the evolutionary history of a stoat will help us to think that provision of water is a relevant positive stimulus, while for a tamarin presence of an aerial predator is a negative one. Animal sentience research helps provide understanding of these mechanisms, both in their operation and their evolution, and thus can help welfare science with right choice of indicators.

https://doi.org/10.1017/psa.2023.10 Published online by Cambridge University Press

3.2. Test for correlated variation in an indicator of the other type

After setting a starting assumption, the second step in validating hidden-target indicators, is to test for correlated variation in an indicator of the other type. This means we measure changes in the set indicator and then look for correlated variation in the indicator we are interested in testing—the test indicator. If we are assuming that variation in the set indicator reflects variation in the target state, then correlation between the set and test indicators should directly reflect correlation between the target and test indicator. This gives us good reason to think that *given the truth of our starting assumption* then the test indicator is a valid indicator of the target.

If the set indicator is a causal indicator and the test indicator an effect, then these tests will ideally take the form of deliberate manipulations on the set indicator, looking to induce associated variation in the test indicator, which stands causally downstream. For example, for tests in animal welfare we can make changes to food availability, or provision of environmental features or even pharmacological interventions, using drugs known (or assumed) to cause changes in welfare-relevant mental states. If the test indicator shows variation alongside the manipulations of the set indicator, this will be presumed to be a result of the changes in the set indicator causing changes in the target, which then cause changes in the test indicator.

If the set indicator is an effect indicator, the tests will be of roughly the same form, but the inferences taken from them will be different. We cannot simply reverse the tests, as the causal direction runs the other ways and manipulations on the effect indicators will not necessarily have any corresponding changes in the causal indicators. Instead, we would still carry out manipulations of the causal indicator and look for correlated changes in the effect indicator. However, given in this case the effect indicator is the set indicator, when we see correlated variation we would then infer the validity of the causal indicator, as our test indicator.

An example of validating a causal indicator might be investigating whether type of handling correlates with welfare changes in sheep. In this case, we would set up tests of different types of handling (human vs. machine) and then use validated effect indicators such as heart rate changes to measure whether a change in welfare is taking place. If a correlation is found, this helps validate the causal indicator. Where we have a cause affecting a target, which in turn affects the indicator, this time the causal link between the target and the effect indicator is based on an assumption, which can then be used to test and validate the link between the causal indicator and the target.

It is important that these tests are done with an indicator of the other type than the set indicator—that is, if the set indicator is causal than the test indicator should be effect, and vice versa—as they stand in different positions in the causal pathway. This is because of differences in validation for the two types of indicator (Bollen and Lennox 1991). While effect indicators can be, in part, validated through measures of correlation with one another, causal indicators *can only be validated through embedding in a model that also contains effect indicators*. This means that testing of causal indicators can only be done using effect indicators. The reverse is not always true. Effect indicators can be validated through testing for correlation with one another.

https://doi.org/10.1017/psa.2023.10 Published online by Cambridge University Press

However, this will only really work when using an effect indicator that is already known to be valid (see step 4 for more on this).

In this stage of assumption-based testing, if both the set and the test indicators are effect indicators, an additional assumption will be required for testing. Although effect indicators will correlate, this is due to being effects of a common cause (the target) rather than a direct causal link. That means that direct intervention on an effect indicator will not necessarily show a change in other effect indicators. Correlated variation will only occur through interventions on the common cause target state, which requires the use of causal indicators. If these causal indicators are not already validated (in which case we are again at step 4), then we are making an *additional* assumption about the relationship between causal indicator and target, that will weaken our tests. Thus all testing at this stage should be of indicators of the other type to that used in the assumption.

As mentioned previously, these tests give us good reason to think that *given the truth of our starting assumption* then the test indicator is a valid indicator of the target. This may seem like a large caveat, if we don't have strong reason to believe in the starting assumption. Our reasons will derive partially from the plausibility described in step 1, and also through the robustness testing that will be described in step 3.

3.3. Repeat tests for the indicator using different assumptions to give robust results

As flagged earlier, there is a weakness so far with the described procedure. That is, that our confidence in our results is only as strong as the starting assumption we have made. This is the role of the third step—to increase our confidence in the results, and thus in the validity of our test indicator, through use of multiple tests, each using different starting assumptions. This type of repeated testing is known as robustness analysis. Animal welfare science often uses a similar process for validation as the one I have outlined so far—to subject animals to a presumed stressor, measure the corresponding effects, and then take these to be valid indicators of stress that can then be used to test for stress under other circumstances (Mason and Mendl 1993). However, what this process misses is the repetition of the tests to test the initial assumption and build robust results.

Robustness is a concept used in much philosophy of science and applied in many different contexts. In a general sense, robustness is something like the property of being "invariant under a multiplicity of independent processes" (Soler 2014, 203). This can apply to a variety of entities and processes, but in this case it applies to observations as the result of various experimental procedures.

Wimsatt (2012) justifies the use of robustness by looking at the impact or errors in different types of reasoning. He describes the traditional scientific method, which aims to establish a small number of fundamental axioms and derive the rest from these. Because there is a small chance of error in any operation in the chain of derivation, long serial chains of reasoning like this will have a much higher chance of error overall. In a serial chain of reasoning, any one step could fail and that will cause a failed result. Small errors in each step multiply, so the more steps there are, the greater the chance of and impact of errors. In Wimsatt's words, "fallible thinkers should avoid long serial chains of thinking" (Wimsatt 2007, 50). By contrast, a parallel or network setup for reasoning will help each strand reinforce the others, as the

https://doi.org/10.1017/psa.2023.10 Published online by Cambridge University Press

chance of error in each one has less chance of impacting the conclusion and this will decrease further with the addition of more lines of evidence. The more steps there are, the more chance of success in the result. We should be more confident in more robust results because of a "no miracles" explanation—it would be a miracle if a variety of independent tests produced the same erroneous result, so the explanation that they are providing an accurate result is more likely (Soler 2014).

The key feature of this sort of analysis is that these lines of evidence are independent. There is a great deal of discussion about what characterizes independence in this context, but the general characterization is one that defines independence in terms of chance of the same types of error occurring. That is, that the differences between the types of tests tries as much as possible to minimize the overlap in the same type of error, so errors are independent and robustness helps build our confidence in the result as described earlier. In this case, what is most important is that the tests rely on *independent background assumptions*. Although all tests will share at least some assumptions, here what matters is that "any *problematic* or *unconfirmed* assumptions should not be shared by the different ways of access" (Eronen 2015, 3969; emphasis in original). If we repeat the tests using different background assumptions, it means that the collective results do not rely on any one assumption in the way that a single test would.

These assumptions should differ in that they use different set indicators, while still testing a single test indicator. For example, we might test an effect indicator of animal welfare first by using a set causal indicator of food quality, which we assume to influence welfare, and then by using the causal indicator of access to social companions. As these two types of causes are different from one another, and the mechanism by which each is thought to affect the target state are different, we would have sufficiently independent assumptions to give robust results. If the tested effect indicator showed the right kind of variation in both cases, we would have good reason to think it is a valid indicator of welfare.

3.4. Use validated indicators as starting point to test others

Once we have used the three steps to validate an indicator for a hidden target, we can repeat for as many indicators as we wish. However, we can also make the process simpler by using the validated indicators to test others ("concurrent validity"—Botreau et al. 2007). The validated indicator would then take the place of the set indicator used in the starting assumption. We can use validated causal indicators as starting points to test effect indicators, and validated effect indicators to test casual indicators. Correlation between a validated indicator and a test indicator tells us they are likely to be mapping onto the same target state, and thus that the test indicator is also valid. Additionally, because of the correlation between effect indicators, as discussed previously, we can also use effect indicators to test one another. Although this will still require assumptions for causal indicators (or use of validated causal indicators), correlation with other validated effect indicators is a strong additional line of evidential support.

As an example, Panksepp (2005) suggests that we could use results from human tests to validate behavioral indicators of welfare in other animals. The suggestion is that we could take neurochemical agents known through self-report to cause changes

https://doi.org/10.1017/psa.2023.10 Published online by Cambridge University Press

in emotional states in humans (e.g., increasing or decreasing joy or sadness). Taking the assumptions that self-report is a reliable enough guide to human experience, and that neurochemical agents are likely to act the same way in other similar brains (i.e., containing similar relevant neural pathways), we can take these causal indicators as valid and then use manipulations in these to test for correlated changes in the effect indicators, such as playful behavior or vocalizations. Where correlated changes are seen, we have good reason to think that these indicators are valid for the changes in welfare.

Because this method does not rely on starting assumptions, but on established validated indicators, it therefore doesn't require the third step of multiple testing for robustness. Our confidence in the validated indicator gives us confidence in the results of the tests. However, in many cases it will still be valuable to run multiple tests. Although the initial testing process may give us confidence in the validity of our tested indicators, it does not give us certainty. Any mistakes in that process would then be amplified if these are then used as the basis for testing others—recall Wimsatt's "chain of reasoning." Running multiple independent tests, using different assumptions or other (independently) validated indicators, gives us increased confidence that there are no such mistakes having an impact on our results, and thus is still a useful step in testing. Our increased confidence in validated indicators as compared to the assumptions of the set indicators might be reflected in the need for fewer lines of testing than we would need initially, but it would usually be advisable to have more than one.

4. Example

There are increasingly many examples in the animal welfare literature of something like this method being used for validating welfare indicators, though without the process being made explicit. Here I will demonstrate how the steps outlined above can map onto the process of validating animal welfare indicators with an example from Briefer et al. (2015), who used a similar method in their promising work in developing indicators to measure both the valence (positive/negative) and strength of welfare in goats.

1. *Make a (plausible) starting* ***assumption*** *relating a causal or effect indicator to the target.*

Goats are placed under differing conditions that are assumed to have positive or negative effects on welfare—access to food or social groups versus being unable to access food and experiencing social isolation. These assumptions—that, for example, access to food improves welfare and seeing but being unable to access it causes reduced welfare—seem fairly plausible and are based on expert knowledge of the animals.

2. *Test for correlated variation in an indicator* ***of the other type.***

The goats are then assessed for changes in various potential effect indicators such as ear position, type of vocalization, and change in heart rate. Those effects that vary

https://doi.org/10.1017/psa.2023.10 Published online by Cambridge University Press

reliably under the different conditions are supported as valid indicators. We have a causal indicator affecting a target, which in turn affects the effect indicator.

3. *Repeat tests for the indicator using* ***different assumptions*** *to give robust results.*

This experiment used two different set indicators for testing, each with its own separate assumption. It is far less likely that the observed effect indicators were indicating some other factor; in the framework, welfare is the most likely link between the food and social conditions.

4. *Use validated indicators as starting point to test others.*

Although not used in this experiment, the indicators that were tested and validated here can form the basis of future testing of both causal and other effect indicators.

5. Conclusion

Subjective animal welfare creates a measurement problem: as it is a hidden target, these subjective states cannot be measured directly and we must instead use indirect indicator measurements such as behavior or physiology. The measurement indicators used must be valid ones—that is, it must be the case that the indicators are measuring the intended target rather than some other target (or nothing at all). This requires a causal relationship between the target and the indicators. This causal relationship can go in either direction—the indicators can either be causes or effects of the target state. These two types of indicators need to be tested against one another for validation. Validating these indicators can be achieved using a four-step approach that requires making some assumptions about the causal links between the target and the indicators and testing these assumptions using multiple independent lines of evidence to increase our confidence in them using robustness analysis. Indicators showing a reliable correlation throughout testing can then be taken to be valid measures of the target state.

References

Beausoleil, N. J., and D. J. Mellor. 2017. "Validating Indicators of Sheep Welfare." In *Achieving Sustainable Production of Sheep*, edited by J. P. C. Greyling, 327–48. *Burleigh Dodds Series in Agricultural Science*. Cambridge: Burleigh Dodds Science Publishing.

Bollen, Kenneth, and Richard Lennox. 1991. "Conventional Wisdom on Measurement: A Structural Equation Perspective." *Psychological Bulletin* 110(2):305–14.

Botreau, R., M. B. M. Bracke, P. Perny, A. Butterworth, J. Capdeville, C. G. Van Reenen, and I. Veissier. 2007. "Aggregation of Measures to Produce an Overall Assessment of Animal Welfare. Part 2: Analysis of Constraints." *Animal* 1(8):1188–97.

Briefer, Elodie F., Federico Tettamanti, and Alan G. McElligott. 2015. "Emotions in Goats: Mapping Physiological, Behavioural and Vocal Profiles." *Animal Behaviour* 99:131–43.

Bringmann, Laura F., and Markus I. Eronen. 2016. "Heating up the Measurement Debate: What Psychologists Can Learn from the History of Physics." *Theory & Psychology* 26(1):27–43.

Browning, Heather. 2020. "If I Could Talk to the Animals: Measuring Subjective Animal Welfare." https://openresearch-repository.anu.edu.au/handle/1885/206204.

Duncan, Ian J. H. 2002. "Poultry Welfare: Science or Subjectivity?" *British Poultry Science* 43(5): 43–52.

https://doi.org/10.1017/psa.2023.10 Published online by Cambridge University Press

Eronen, Markus I. 2015. "Robustness and Reality." *Synthese* 192(12): 961–77.
Lindenmayer, David B., and Gene E. Likens. 2011. "Direct Measurement versus Surrogate Indicator Species for Evaluating Environmental Change and Biodiversity Loss." *Ecosystems* 14(1):47–59.
Markus, Keith A., and Denny Borsboom. 2013. *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. New York: Routledge.
Mason, Georgia, and Michael Mendl. 1993. "Why Is There No Simple Way of Measuring Animal Welfare?" *Animal Welfare* 2(4): 01–19.
Mellor, D. J., and N. J. Beausoleil. 2015. "Extending the 'Five Domains' Model for Animal Welfare Assessment to Incorporate Positive Welfare States." *Animal Welfare* 24(3):241–53.
Panksepp, Jaak. 2005. "Affective Consciousness: Core Emotional Feelings in Animals and Humans." *Consciousness and Cognition* 14(1):30–80.
Schickore, Jutta, and Klodian Coko. 2013. "Using Multiple Means of Determination." *International Studies in the Philosophy of Science* 27(3):295–313.
Soler, Léna. 2014. "Against Robustness? Strategies to Support the Reliability of Scientific Results." *International Studies in the Philosophy of Science* 28(2):203–15.
Wimsatt, William C. 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge: Harvard University Press.
Wimsatt, William C. 2012. "Robustness, Reliability, and Overdetermination (1981)." In *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science*, edited by Léna Soler, Emiliano Trizio, Thomas Nickles, and William C. Wimsatt, 61–87. Dordrecht: Springer.

Cite this article: Browning, Heather. 2023. "Validating Indicators of Subjective Animal Welfare." *Philosophy of Science* 90 (5):1255–1264. https://doi.org/10.1017/psa.2023.10