ARTICLE IN PRESS

Studies in History and Philosophy of Science xxx (2017) 1–10

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

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

# The appeal to robustness in measurement practice

Alessandra Basso

*Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences, Department of Political and Economic Studies/Philosophy, University of Helsinki, PL 24 00014 Helsinki, Finland*

ARTICLE INFO

*Article history:*
Received 3 December 2015
Received in revised form
26 May 2016
Available online xxx

*Keywords:*
Robustness argument
Measurement reliability
Measurement in the social sciences

ABSTRACT

This paper distinguishes between two arguments based on measurement robustness and defends the epistemic value of robustness for the assessment of measurement reliability. I argue that the appeal to measurement robustness in the assessment of measurement is based on a different inferential pattern and is not exposed to the same objections as the no-coincidence argument which is commonly associated with the use of robustness to corroborate individual results. This investigation sheds light on the precise meaning of reliability that emerges from measurement assessment practice. In addition, by arguing that the measurement assessment robustness argument has similar characteristics across the physical, social and behavioural sciences, I defend the idea that there is continuity in the notion of measurement reliability across sciences.



## 1. Introduction

In the philosophy of science, the notion of robustness refers generally to situations in which something is stable under variations of something else. The epistemic import of robustness has been thoroughly debated in the literature, in relation to various scientific practices and different epistemic tasks (Cartwright, 1991; Weisberg, 2006; Woodward, 2006; Stegenga, 2009, 2012; Kuorikoski, Lehtinen, & Marchionni, 2010, 2012; Alexandrova & Odenbaugh, 2011; Hudson, 2013; just to mention a few).

Measurement robustness is achieved in the case that a number of different procedures yield closely similar results for a certain quantity under measurement (see Woodward, 2006). The employment of multiple procedures for corroborating measurement results has most commonly been interpreted as being based on the following inference, that here I call the no-coincidence argument: if (i) multiple procedures converge on compatible results, and; (ii) the procedures are independent in some relevant sense, then the confidence in the correctness of the robust result increases. The correctness of the result provides a most plausible explanation of the convergence. In contrast, it would be an implausible coincidence that multiple independent procedures converged on the same results if the results are wrong or the procedures highly inaccurate. A canonical example is Perrin's derivation of Avogadro's number by means of multiple methods, which has been understood as a case in which a scientific result is supported by showing that it is robust across multiple independent methods of derivation (Cartwright, 1983, 1991; Salmon, 1984; Woodward, 2006).[1,2]

The no-coincidence argument is exposed to various objections that are well-rehearsed in the literature. In particular, here I focus on two main objections: first, that the presumption of independence is too ambiguous and, in its most plausible formulations, hardly ever met, and second that the convergence of reliable procedures is ultimately uninformative. These objections appear to be especially relevant for the context of measurement, and therefore, if the rationale behind the appeal to robustness in measurement practice were always the one described by the no-coincidence argument, then, in many cases, measurement robustness would be of dubious epistemic import.

In this paper, however, I argue that there is another, often overlooked argument based on measurement robustness, which has a different inferential structure and is not exposed to the same objections. According to this argument, robustness allows one to evaluate and improve the reliability of measurement procedures. The epistemic value of measurement robustness in this context

[1] Drawing on this seminal example, some authors seem to believe that the use of this argument in the context of measurement is more straightforwardly confirmational than in other contexts – particularly in comparison to derivational or inferential robustness (Cartwright, 1991; Woodward, 2000).

[2] Based on a no-miracle argument, Perrin's multiple derivation of the Avogadro's number was also taken as evidence of the existence of atoms (see e.g. Salmon, 1984, pp. 214–227). This paper, however, is not intended to discuss this argument for scientific realism.

*E-mail address:* alessandra.basso@helsinki.fi.

http://dx.doi.org/10.1016/j.shpsa.2017.02.001
0039-3681/

Please cite this article in press as: Basso, A., The appeal to robustness in measurement practice, Studies in History and Philosophy of Science (2017), http://dx.doi.org/10.1016/j.shpsa.2017.02.001

stems from the fit between actual and expected convergence, rather than from the simple convergence of measurement results. More precisely, the 'measurement assessment' robustness argument is based on expectations about how the procedures should converge given their different flaws and weaknesses, that is, the different ways in which they can fail to measure the quantity under measurement as defined. The investigation of the fit between actual and expected convergence allows one to evaluate the coordination between the definition of the quantity under measurement and the measurement procedure. This assessment practice, therefore, reveals a specific meaning of measurement reliability, which is related to how well the outcomes approximate the idealized definition, and thereby gives an indication of how well a procedure measures what it is meant to measure.

The appeal to robustness within measurement assessment practice has received less attention from the philosophical literature in comparison to its use for corroborating individual results, especially when it comes to measurement in the social sciences. This paper illustrates these two arguments by means of examples both from the physical and from the social and behavioural sciences, and argues that the central features of these arguments recur across various sciences. This investigation prompts an exploration of some similarities and differences between measurement practices in distinct scientific contexts. In particular, the similarities in the appeal to robustness in the assessment of measurement suggest that there is continuity between natural, social and behavioural sciences in the meaning of measurement reliability and its assessment methods.

The paper proceeds as follows. In Section 2, I describe the no-coincidence argument for corroborating individual measurement results and discuss the major objections raised against this kind of argument. In Section 3, I illustrate the appeal to robustness in the assessment of measurement: I provide examples both from the natural sciences and from the social and behavioural sciences and identify the similarities and differences between the respective disciplinary contexts. In Section 4, I reconstruct the inferential pattern behind the appeal to robustness in the assessment of measurement and discuss how it differs from the no-coincidence argument. I argue that the two arguments are not exposed to the same objections. Section 5 summarizes and concludes.

## 2. The appeal to robustness for corroborating measurement results

The comparison of different procedures to cross check individual measurement results is common practice both in the natural and in the social sciences. A canonical example in the philosophical literature is Perrin's derivation of Avogadro's number by means of thirteen different methods. The significant consistency with which Perrin was able to derive his result using a number of reliable and independent methods is taken to provide strong reasons for the correctness of the robust result (Cartwright, 1983, pp. 82–85, 1991; Salmon, 1984, pp. 214–227; Woodward, 2006; cf. also van Fraassen, 2009; Hudson, 2013, pp. 103–138).

In the social sciences, in psychology and in psychiatry, it is also common practice to support the results of an inquiry by showing that they are robust across various measurement procedures. In behavioural economics, for instance, a study aimed at investigating the effects of organizational incentives on firm performance or labour productivity can use different measures of firm performance or labour productivity in order to cross check the results. Similarly, in social psychology, a study of the correlation between individuals' embarrassability and their pro-sociality (that is, the individual's propensity to care about others), would employ different measures of pro-sociality including both behavioural and psychological measures (see e.g. Feinberg, Willer, & Keltner, 2011). In these studies, only the results that are robust across the different measures are endorsed, whereas results that are not robust are considered disputable. By endorsing only the robust results, scientists aim to ensure that their conclusions are not artefacts of the measurement procedures they used.

A similar way of appealing to measurement robustness can also be observed across studies, when drawing robust results within a field of inquiry. For instance, the well accepted claim that there is comorbidity between substance abuse and mental disorders is supported by appeal to a variety of empirical studies investigating this phenomenon in relation to different addictive substances, distinct mental disorders and different populations, and which use multiple measures for evaluating substance addiction and mental health (for a review, see Jané-llopis & Matytsina, 2006). By showing that the correlation holds under alternative measurement procedures and across different empirical studies, scientists support their findings and avert criticism.

Let us consider an example in more detail. In behavioural economics, the alleged detection of the endowment effect has been supported by appealing to the convergence of results in a number of different surveys and experiments. The endowment effect refers to the claim that people tend to ascribe additional value to things merely because they own them. This claim has been supported by appealing to a remarkable variety of surveys and experiments in which it is observed that people tend to pay more to retain something they own than to obtain something. More precisely, the robustness of the observation that the sellers' willingness to accept (WTA) is higher than the buyers' willingness to pay (WTP) has been taken as deriving from a stable feature of people's preferences, the endowment effect.

In one of the most influential studies supporting the endowment effect, participants were given a mug and then offered to sell it or trade it for an equally priced alternative (Kahneman, Knetsch, & Thaler, 1990). Kahneman et al. found that, once their ownership of the mug had been established, the amount that participants required as compensation for the mug was more than twice as high as the amount they were willing to pay to acquire the mug. Another example of the endowment effect is provided by Carmon and Ariely's (2000) survey, which showed that the price at which participants were prepared, hypothetically, to sell National Collegiate Athletic Association basketball tournament tickets was 14 times higher than the price that they were prepared, hypothetically, to pay for those same tickets. In addition to these examples, the WTP-WTA gap has been observed in studies drawing on a wide range of goods, with different populations, experimental designs and elicitation techniques. For instance, the WTP-WTA gap has been observed in studies involving ordinary market goods such as mugs and chocolates, environmental amenities such as unpolluted air or uncontaminated food, or rights like hunting licences and others. The employed experimental designs included various settings of experimental markets as well as hypothetical markets (surveys), and a number of elicitation techniques, such as incentive-compatible open-ended question, payment card and iterated closed-ended question. The endowment effect has also been investigated across different experimental subjects including both student and non-student populations, children, great apes and New World monkeys (for reviews, see Hoffman & Spitzer, 1993; Horowitz & McConnell, 2002).

As a result, the endowment effect is considered by some as one of the most robust findings in the psychology of decision making (Knetsch, Tang, & Thaler, 2001). The convergence of different procedures for detecting the WTP-WTA gap has been taken as a sign that the endowment effect is a stable feature of people's preferences, rather than a mere artefact of the features of the