Do we measure overconfidence? A closer look at the interval production task[☆]

Ferdinand Langnickel[a], Stefan Zeisberger[b],[*]

[a] Department of Banking and Finance, University of Zurich, Plattenstrasse 32, 8032 Zurich, Switzerland
[b] College of Business, Stony Brook University, Stony Brook, NY 11794, United States

ABSTRACT

The most common test for overconfidence in the form of miscalibration—the interval production task (IP)—is based on the assumption that people internalize requested confidence levels. We demonstrate experimentally that decision makers' perceived confidence is, however, unaffected by variations in the requested confidence level. In addition, we find large heterogeneity in perceived confidence that the traditional IP measure fails to account for. We show that the alternative measure based on decision makers' perceived confidence by contrast yields coherent, moderate overconfidence levels. Our evidence suggests that the consistency of the two measures is limited and that they are related to different individual characteristics.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Numerous studies have demonstrated that overconfidence can lead to suboptimal decisions in a variety of domains (Camerer and Lovallo, 1999; Johnson and Tierney, 2011; Malmendier and Tate, 2005; Odean, 1998). The most robust documentation of overconfidence is that people overestimate the precision of their own judgment (for a review, see Lichtenstein et al., 1982; Moore et al., 2014). This form of overconfidence, called miscalibration, overprecision, or judgmental overconfidence, induces people to rely too much on their own (biased) judgment and thus is an integral part of biased decision making in general (Bazerman and Moore, 2013).[1] The mainly used test for it is the interval production task (IP).[2] In the IP, decision makers are asked to provide lower and upper bound estimates (intervals) for a set of questions like "How long is the Nile river?". Subjects are instructed to state intervals such that their own confidence, that the true, unknown value falls

---

[1] In the following, we will use only the term overconfidence to refer to judgmental miscalibration, unless stated otherwise.
[2] See, e.g. Alpert and Raiffa (1982), Ben-David et al. (2013), Biais et al. (2005), Russo and Schoemaker (1992).

between these stated bounds, equals a confidence level that is requested by the experimenter, for example 90%. On average, the ratio of true values that fall into decision makers' interval estimates, the "hit rate", should correspond to the requested confidence level (in this case 90%). However, commonly people are found to have much lower hit rates, so that they are classified as overconfident (Alpert and Raiffa, 1982; Russo and Schoemaker, 1992).

This paper aims to critically assess the suitability of the IP to measure overconfidence. Recent studies have already suggested that the IP measure may not function as presumed. In particular, it has been shown that groups with different requested confidence levels achieve the same average hit rate because they do not adjust the width of their interval estimates (Teigen and Jorgensen, 2005).[3] In addition, when estimating their own number of hits ("frequency judgments") people tend to be only conservatively confident in their own intervals compared to the typically high level of requested confidence (Cesarini et al., 2006; Winman et al., 2004). What remains unsolved, however, is the actual effect of the requested confidence level on the *perceived* level of confidence. Will people, for example, still provide conservative frequency judgments if the requested confidence level is low?

An experimental economist who studies the relation of overconfidence and economic behavior nonetheless can be agnostic about these points of critique as they address the aggregate, not the individual level. Therefore, it is not surprising that the IP has been used in numerous experimental studies to elicit overconfidence at the individual level and link it to economic behavior (for most recent studies, see Ackert et al., 2015; Ben-David et al., 2013; Fellner-Röhling and Krügel, 2014; Herz et al., 2014), but the results are still surprisingly inconclusive overall.

Against this background, we address three open issues to help understanding the measurement of aggregate and individual overconfidence in the IP paradigm. First, one might wonder why decision makers are unable to adapt their intervals widths in the IP. To provide an answer we take a step back and test whether the requested confidence level has an effect on people's *perceived* level of confidence in the first place. Second, given the often unsuccessful experimental attempts to link overconfidence and economic behavior at the individual level, another important open question is whether the use of frequency judgments changes the relative ranking in overconfidence compared to the classic IP measure. Lastly, the two overconfidence measures may be explained by individual "background" characteristics such as cognitive abilities or the aversion of wide intervals which possibly explains some experimental results.

The IP overconfidence measure rests on the assumption that people adopt the requested confidence level and adjust their interval estimates accordingly. Using the classical IP paradigm we run an experiment in which we elicit *perceived confidence levels* in the form of frequency judgments, i.e. ex-post estimates of the number of hits in the IP, and we vary the degree of requested confidence. In two independent surveys in Switzerland and the U.S., we employ a between-subject design in which participants are randomly assigned to one of three treatments with requested confidence levels of 30%, 60% or 90%. We confirm weaknesses of the IP measure presented in Teigen and Jorgensen (2005) and show that decision makers not even adjust their frequency judgments to different levels of requested confidence. Using decision makers' frequency judgments, we find evidence that people respond to an individual confidence level that is unaffected by the requested confidence level. In all treatments, we observe large variations in individual confidence that the IP measure does not take into account. As a consequence, the consistency of the two overconfidence measures is limited. People might appear very overconfident in the IP simply because their true confidence level is overestimated by the requested level and vice versa. We conclude that the IP as it is currently used, i.e. comparing hit rates with requested confidence levels, has major shortcomings for measuring overconfidence at both the aggregate and individual level. As an alternative, we propose to use people's frequency judgments to measure overconfidence.

We would like to mention that we are not the first to question the reliability of measuring overconfidence with the IP on the aggregate level. Next to the aforementioned studies (Teigen and Jorgensen, 2005; Cesarini et al., 2006; Winman et al., 2004), many previous studies have raised specific methodological concerns with consequences for the aggregate level of overconfidence. It has been shown, for example, that the alternative two-choice question format (e.g., Koriat et al., 1980) yields lower levels of overconfidence than the IP (Klayman et al., 1999). Similarly, exclusion instructions that induce people to think about the values that lie outside of their provided intervals in the IP yield smaller levels of overconfidence than inclusion questions (Soll and Klayman, 2004; Teigen and Jorgensen, 2005; Yaniv and Schul, 1997). Interestingly, other forms of overconfidence such as better-than-average beliefs (Svenson, 1981) or the illusion of control (Langer, 1975) were found to be inconsistent with the IP measure (Deaves et al., 2008; Glaser and Weber, 2007; Hilton et al., 2011; Menkhoff et al., 2006; Moore and Healy, 2008). Moreover, the level of overconfidence has been found to vary with the difficulty (Lichtenstein et al., 1982) and domain (Klayman et al., 1999) of the question set. Other studies attribute overconfidence to measurement errors and sampling effects of the question set that impair the validity of otherwise well-working cues (Gigerenzer et al., 1991; Juslin, 1993, 1994; Soll, 1996). Our work contributes to the understanding of how appropriate the IP measure is on the aggregate level.

Additionally, our findings can help to explain previous puzzling results on the individual level. While the IP has been found to predict some behavior like innovative activity, risk-taking and ordering decisions of managers (Ben-David et al., 2013; Herz et al., 2014; Ren and Croson, 2013), a wide range of studies is not able to explain economic behavior with the IP overconfidence measure. In a recent study, for example, Ackert et al. (2015) find that investors who diversify more in an

---

[3] These findings are based on between-subject variation of the requested confidence level. In response to within-subject variation of the requested confidence level decision makers have been shown to adjust their interval estimates (see, e.g. Alpert and Raiffa, 1982; Budescu and Du, 2007).