Open Peer Review on Qeios

Trust is the best policy. Game theoretical analysis of bias in elicitation procedures in linguistics

Hubert Kowalewski[1]

1 Maria Curie-Skłodowska University (MCSU)

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

Putative bias of consultants is often quoted as a major obstacle in elicitation procedures in linguistics. A frequently recommended solution to the problem is making sure that consultants are "naive," i.e. they do not possess expert knowledge in linguistics or at least they are ignorant of research hypotheses and expected results of the elicitation. If "naiveness" is indeed the cornerstone of reliable elicitation procedure in linguistics, linguists' intuitive judgments about linguistic phenomena are inherently and unavoidably flawed. I propose, however, that bias is not as major a problem as it is frequently believed, not even when consultants are linguists themselves. The analyses within the mathematical frameworks of game theory and decision theory demonstrate that even when some consultants are display clear bias of some sort, consistent acceptance of all intuitive judgments is the best strategy for harvesting reliable data. While universal acceptance does not safeguard researchers against data distorted by bias, it yields legitimate data in a clear majority of cases under almost all reasonable assumptions about the distribution of bias in the population of consultants.

1. Introduction

The usefulness of intuitive judgments as data for linguistic theorizing is often questioned because of putative bias of consultants, especially in the case of "self-elicitation," when the linguist uses their own intuitions. Intuitive judgments have defenders: Noam Chomsky (1965), Eugenio Coseriu (1991), Esa Itkonen (2005, 2008), Raymond Gibbs (2006), Ronald Langacker (2008) are among authors who, in one way or another, endorse intuitive judgments as legitimate data in linguistics, albeit with certain limitations. Nonetheless, the conviction that bias is somehow inherent in any elicitation procedure and it inevitably compromises the reliability of the speakers' judgments is very much alive within the community of linguists. This article aims to undercut this conviction that bias is always a major obstacle during elicitation. Of course, I do not lightheartedly assume that consultants are never biased; on the contrary, the analyses in the following sections assume that bias affects a significant portion of consultants. I do deny, however, that bias is always detrimental and biased consultants invariably offer corrupted judgments about linguistic data under investigation.

The main part of the article provides mathematical analyses of the elicitation procedure in terms of game and decision

theories. The theories are formal frameworks used for modeling decisions made during interactions of agents (game theory) or in “non-interactive” situations involving possible states of affairs (decision theory). In general, game- and decision-theoretical analyses consist in weighing various strategies available for decision-makers against expected outcomes in order to maximizes gains or minimizes losses. The analyses conducted in the article demonstrate that the negative effect of consultants’ putative bias is less severe than often claimed and that consistent acceptance of all intuitive judgments is linguist’s optimal strategy in a clear majority of cases. Sections 3 and 4 present analyses based on somewhat simplified and idealized mathematical models. Since every model is only as good as the degree to which is approximates real life, in Section 5 I will revise the key assumptions behind the analysis in order to increase its compatibility with real-life situations.

Game theory has found many applications in social sciences, most notably in economics, but in recent years it has found its way into linguistics as well. The applications of this framework, especially in the field of semantics and pragmatics, are showcased in the volumes *Game Theory and Linguistic Meaning* (Pietarinen 2007) and *Meaningful Games: Exploring Language with Game Theory* (Clark 2011). Gerhard Jaeger uses the framework to analyze the process of interpreting linguistic meanings (Jaeger 2008). My article, in turn, was inspired by Francesco Guala’s analysis of bias in social sciences, most notably sociology and economics (cf. Guala 2016).

There are several reasons why one may be skeptical about the usefulness of intuitive judgments in linguistics. Due to their complexity, limitations of space and the scope of this article, it is impossible to discuss them in detail, but a tentative survey may be helpful in spelling out the aim of this article more clearly. Michael Halliday and Christian Matthiessen, drawing on Halliday’s earlier work (cf. Halliday, McIntosh, and Strevens 1964), are skeptical about intuitive judgments, because “[what] people actually say is very different from what they think they say; and even more different from what they think they ought to say” (Halliday and Matthiessen 2004, 34). I grant that there is a discrepancy between speakers’ actual language and speakers’ judgments about their language use, but this argument only shows that if one has already decided to study actual language use, speakers’ judgments cannot be treated as reliable source of data. It does not demonstrate, however, that one should choose to study actual language use rather than, for example, linguistic competence of speakers, which is never fully reflected by real-life language use. If one is more interested in the competence after all, one could point out the discrepancy between competence and performance à la Noam Chomsky (cf. 1965) or between *langue* and *parole* à la Ferdinand de Saussure (cf. 1966 [1916]) to vindicate speakers’ judgments as a valuable source of data. Yet the discussion about legitimate sources of data is largely independent of the discussion on ways to handle putative bias of intuitive judgments. I will therefore leave the former aside and assume that some linguists do take intuitions as legitimate data, but may still worry about consultants’ bias.

One may also question speakers’ ability to produce valuable intuitions about linguistic expressions. The objection effectively severs the connection between speaker’s ability to use language in “natural” usage events and to evaluate linguistic expressions in “artificial” contexts of judgment elicitation. The objections merits a more complete treatment than I can provide in this article, but the short answer is that the ability to produce judgments is simply an inherent part of linguistic competence. The competence can be thought of as the so-called multi-track disposition as analyzed by Gilbert Ryle (2009 [1949]). Multi-track dispositions produce many different behaviors in many different circumstances and all of

the behaviors are “legitimate” manifestation of the disposition. In other words, to be a speaker of a language is not only to be able to produce and understand linguistic expressions, but also to recognize well-formed and malformed expressions in that language, to correct the malformed expressions, to translate an expression into a different language known to the speaker, etc. Thus, in the following part of the article it is assumed that speakers are capable of forming intuitive judgments simply by virtue of being speakers of a given language.

Some linguists, especially those of inductionist proclivities, point out that difficulties with achieving consistent intuitive judgments across consultants renders intuitions virtually useless in practice (the problem is acknowledged, among others, by Wasow and Arnold (2005) and discussed extensively by Schütze (2016)). This issue is once again independent of the problem of putative bias, since inconsistency of intuitive judgment neither entails, nor precludes consultants’ bias. Instead, the problem seems to result from the tension between linguist’s desire to achieve valuable generalizations about language and unavoidable idiolectic differences across speakers. There are no perfect and universal ways of resolving the tension, although reasonable and workable solutions may be found for specific research problems.

Yet another objection arises from notorious confusion between broadly defined introspection and intuitive judgments. As Kurt Danziger (1980) shows, the modern disreputation of introspectionism is not entirely deserved and it is exacerbated by prejudiced historical accounts. Nonetheless, it would be indeed hard to argue that the mechanisms of the mind are fully revealed in acts of introspective insight. Yet despite the lack of introspective insight into the psychological and cognitive machinery underlying linguistic faculty, one may still hold that speakers are able to deliver various kinds of reports about linguistic expressions. A speaker of English may entertain and report the impression that the sentence *The cat is on the mat* is a well-formed sentence in English without knowing what makes the sentence well-formed and without explicit declarative knowledge of the mechanisms of their mental grammar.[1] The justified skepticism about the epistemic powers of introspection leads some authors to overgeneralized skepticism about all sorts of introspective insight, including basic access to subjective impressions resulting in intuitive judgments.[2] It is worth mentioning, however, that other authors successfully avoid this confusion; for example, Wasow and Arnold (2005) draw a distinction between primary intuitions (judgments about expressions) and secondary intuitions (judgments meant to explain why expressions have certain properties). Following the authors’ nomenclature, the article discusses putative bias in primary intuitions and the discussion and remains silent about putative bias in secondary intuitions.

2. Decision theory and game theory

Game theory is a mathematical study of interactions between decision-makers. Decision theory is a study of decision-making in non-interactive situations, when the agent chooses an action depending on possible states of affairs, rather than on decisions of another agent. Mathematical methods in both frameworks are very similar. Typically, the decision-makers are assumed to be rational, i.e. they have certain goals and choose their strategies to maximize gains, called “payoffs” in the game-theoretical nomenclature. In two-player games the available strategies can be conveniently presented in a payoff matrix, which summarizes players’ “gains” and “losses” relative to the strategies selected by each of them. Decision theory matrices feature only the payoffs of the decision-maker relative to certain states of affairs. Table 1 is

the (game-theoretical) payoff matrix for a driving game, in which two cars drive against each other and the drivers can choose to turn either left or right. When both drivers choose to turn either left, or right, they avoid the crash and they both win. Otherwise, the cars crash and both drivers lose. The strategies available for the first driver are presented in the first (gray) column, while the strategies available for the second driver are presented in the first (gray) row. For the sake of simplicity, I represent the payoffs numerically as numbers within the range from -1 to 1, 1 being the maximum payoff.

Table 1. The payoff matrix in the driving game

	turn left	**turn right**
turn left	**1, 1**	-1, -1
turn right	-1, -1	**1, 1**

A payoff matrix makes is easier to find the optimal strategy for each player. In the driving game, there are two optimal strategies for each player: it is optimal for Driver 1 (in the column) to turn left when the Driver 2 turns left or to turn right when the Driver 2 turns right. Analogically, the same holds true for the optimal strategy of Driver 2. The driving game has perfect information, since each player knows the other player's available and optimal strategies. The players' optimal strategy is the so-called **Nash equilibrium**. Rational players gain nothing from changing their optimal strategies, so they have no incentive to do so. In a two-player game, a convenient rule of thumb for finding the Nash equilibrium is to look for a cell, where the left-hand value is the maximum out of all left-hand values in its column and the right-hand value is the maximum out of all right-hand values in its row.

In Table 1, the Nash equilibria can be identified by applying the above-mentioned rule of thumb and are marked in bold font. Since in the driving game there are two equilibria with equal payoff values for both players, the crucial factor is whether each player knows which strategy is selected by the other player. For example, if Driver 1 knows that Driver 2 is going to turn left, Driver's 1 optimal strategy is to turn left. Of course, if Driver 1 does not know what Driver 2 is going to do, her best strategy is to assume that there is a 50% probability for Driver 2 to turn left and 50% probability to turn right. In effect, Driver 1 has to choose the strategy at random. Strategies involving random choice between several available strategies with some probability of choice ascribed to each of the options are termed **mixed strategies**. Thus, the driving game has three Nash equilibria: two pure equilibria marked in bold in Table 1 and one mixed strategy equilibrium involving random choice between "turn left" and "turn right" with the probability of 50% each (the above-mentioned rule of thumb does not reveal mixed- strategy Nash equilibria, which need to be calculated differently).

3. The elicitation game

The procedure of eliciting intuitive judgments from a consultant, discussed in this section, can be modeled mathematically

as a complex game subsuming two subgames between the linguist and the consultant. Section 4 discusses the procedure of self-elicitation, when linguists use their own intuitive judgments. Such elicitation procedures are modeled more conveniently in terms of decision theory, since it would be hard to argue that during self-elicitation the linguist interacts with another player distinct from themselves. Arguably, self-elicitation is more controversial among linguists, due to the conviction that the researcher are inherently biased, which prevents them from delivering reliable judgments.

In both types of elicitation procedures, the linguist has two strategies to choose: they may either accept the judgment as reliable or reject it as unreliable. When another person is used as a consultant, the consultant may choose to deliver a veridical judgment, i.e. report their actual impression to best of their abilities, or deliver a falsidical judgment, i.e. report the opposite of their actual impression. The consultant may have various reasons to deliver either veridical, or falsidical judgment depending on their overall goals. I will consider three scenarios: elicitation from an unbiased consultant, a positively biased consultant, and a negatively biased consultant.

3.1. Unbiased consultants

An unbiased consultant has no stake in the outcome of the elicitation procedure and they will not try to please or mislead the linguist. The goal of an unbiased consultant is simply to deliver a veridical judgment, so obviously their optimal strategy is to report veridically. In game-theoretical terms, delivering a veridical judgment is always the winning strategy and delivering a falsidical judgment is always the losing strategy. The payoff matrix for this game is shown in Table 2, where linguist's strategies are in the first column and consultant's strategies in the first row.

Table 2. Payoff matrix for unbiased consultant elicitation

	veridical	**falsidical**
accept	**1, 1**	-1, -1
reject	-1, 1	1, -1

Unsurprisingly, the conclusion from Table 2 is that the Nash equilibrium in the unbiased consultant scenario is for the consultant to produce a veridical judgment and for the linguist accept it. A distrustful linguist would maximize their payoff if they reject the falsidical judgment, but since producing such a judgment is not in the best interest of the consultant, it cannot be expected that the strategy will by selected (therefore, the lower-left cell in Table 1 is not a Nash equilibrium). I do not expect this analysis to be particularly controversial, but it is also hardly revealing. Obviously, If all consultants were unbiased, the objections against intuitive judgments would not get off the ground in the first place. The real problem is that some consultants may be, in fact, biased in some way. In principle, a consultant may display positive or negative bias.

3.2. Positively biased consultants

A positively biased consultant is overly conforming and eager to please the researcher; such consultants try to deliver a judgment that (to the best of their knowledge) the linguist wants to receive. The strategies available for a biased consultant are different from the ones available to the unbiased person since a biased person's goal is not to offer a veridical judgment, but to offer the expected judgment, regardless of whether this judgment is indeed veridical. Typically, however, elicitation procedures are designed in such a way that consultants do not know which judgment is expected. Thus, the situation is characterized by incomplete information, when the consultant needs to choose the strategy in the absence of critical information about researcher's expectations. The standard procedure for analyzing games with incomplete information about one of the players is to reduce them to the so called **Bayesian game**, where all possible types of the "unknown" player are considered. The types are assigned probabilities reflecting the belief of one player about the likelihood of the "unknown" player instantiating a given type. This maneuver transforms a game with incomplete information (where the goals of the linguist are unknown) into a game with imperfect information (where the goals of the linguist are known with some degree of probability). Since games with imperfect information have a mixed-strategy Nash equilibrium, it is now possible to pursue the optimal strategy.

Let us assume that the researcher expects that the expression under investigation has the property P (e.g. the expression is grammatical). If a positively biased consultant reports that the expression does have P, the consultant "wins" by complying with researcher's expectations and the researcher "wins" by having their expectations satisfied. Analogically, if the researcher expects the expression to lack P (e.g. the expression is ungrammatical) and the consultant reports that the expression lacks P indeed, both the consultant and the researcher "win." In any other situation both players lose: the consultant by confounding the researcher's expectations and the researcher by having their expectations confounded. Thus, the game results in **the positive consultant's dilemma** (PCD) summarized Table 3 (where the first column represents linguist's expectations about the expression and the first row possible judgments by the consultant). Incidentally, the positive consultant's dilemma has the mathematical structure of the already mentioned driving game (see Table 1).

Table 3. Positive consultant's dilemma

	P	**¬P**
P	**1, 1**	-1, -1
¬P	-1, -1	**1, 1**

In principle, the game has two pure Nash equilibria – to report P when the linguist expects P and to report ¬P otherwise – but since the consultant has incomplete information about researcher's expectation, they cannot make an unequivocal

decision about which pure Nash equilibrium should be chosen. In order to determine the best strategy, the consultant may convert the game into a Bayesian game by construing to "types" of linguist: one that expects P and the other that expects ¬P. The next step is to assign probabilities to the types. If the consultant is genuinely in the dark about the linguist's expectations, the consultant has no reason to assign higher probability to any of the types and the best move is to ascribe equal probability (50%) to each type. Even without sophisticated mathematical machinery it is now apparent that the consultant wins in 50% of cases regardless of whether they report P or ¬P, which is the consultant's mixed Nash equilibrium.[3]

This intuitive conclusion is consistent with more rigorous mathematical approach. To find the optimal consultant's strategy in a Bayesian game, it is necessary to calculate the expected consultant's payoffs for each "type" of linguist (i.e. for each linguist's expectation). The payoffs can be calculated with the formula below, where $payoff_X$ stands for payoffs after reporting X and $prob_X$ is the probability that the linguists expects X.

$$prob_X \times payoff_X + \left(1 - prob_X\right) \times payoff_{\neg X}$$

If the linguist expects P, $payoff_P$ and $payoff_{\neg P}$ equal 1 and -1 respectively. If the consultant has no reason to believe that either linguist's expectation is more likely, the probability of linguist expecting P equals 50% (represented as 0.5 for convenience). Hence:

$$0.5 - (1 - 0.5) = 0$$

If the linguist expects ¬P, the consultant's payoff after reporting P is also 0:

$$-0.5 + (1 - 0.5) = 0$$

This simply means that consultant's wins and losses after reporting P cancel each other out given the linguist expects P with 50% probability. Unsurprisingly, the situation is symmetric when the consultant reports ¬P: given 50% probability of each type of linguist, the consultant's payoffs are 0. Since neither of the strategies "report P" and "report ¬P" yields a better result, the Nash equilibrium is to produce P and ¬P with 50% probability each.

What about the linguist's payoffs? A crucial point to note is that consultant's positive bias does not mean the report is necessarily different from the report delivered by an unbiased consultant. If a positively biased consultant believes the sentence (1) is expected to be grammatical and judges that it is grammatical to please the researcher, it is still possible

that the consultant would also judge (1) as grammatical in “neutral” circumstances, outside elicitation procedure when the bias did not influence their judgment.

(1) The dog chased a cat.

In such a case, bias stabilizes the judgment which the consultant would produce anyway and there is no reason for rejecting such a judgment as distorted or unreliable. Admittedly, a more worrying situations is when the positive bias alters the judgment and the linguist should reject it. Note that a positively biased consultant does not provide their judgment on the basis of their heartfelt impressions, but on the basis of their beliefs about linguists expectations. Therefore, the judgment delivered may be stabilized or altered regardless of any strategic considerations on the part of the consultant. Obviously, there is no way for the linguist to decide whether consultant’s judgment is stabilized or altered, so it is now the linguist that needs to make the decision under uncertainty. The situation could be termed **the linguist’s dilemma** (LD), summarized in Table 4. In LD the goals and strategies of the consultant in PCD are irrelevant; the linguist is only interested in accepting a stabilized judgment and rejecting the altered one. Thus, the situation is more adequately modeled within the framework of decision theory rather than game theory, since the linguist is simply confronted with two possibilities – the judgment being either stabilized or altered – and needs to make the optimal decision. Thus, in Table 4, the first row does not present strategies eligible by the consultant, but possible state of affairs and the values in the cells represent payoffs expected by the linguist when their decisions are weighed against respective states of affairs.

Table 4. Linguist’s dilemma

	stabilized	**altered**
accept	**1**	-1
reject	-1	**1**

Unsurprisingly, the linguist maximizes their payoff by either accepting a stabilized judgment of rejecting the altered one. Since the linguist does not know whether the judgment is stabilized or not, the best they can do is to resort to **the Laplace criterion**, which is to ascribe equal probability to all options. Under the criterion, an elicitation from a positively biased consultant yields 50% of stabilized and 50% of altered judgments. Consequently, both consistent acceptance and consistent rejection of judgments is an optimal strategy in 50% of cases.

3.3. Negatively biased consultants

Negatively biased consultants are the opposite of the positively biased ones: they try to defeat linguist’s expectations

rather than to satisfy them. This kind of bias does not seem to particularly worry linguists and literature on methodology of linguistics does not typically warn against deceitful consultants plotting to sabotage elicitation with their dishonest reports. Nonetheless, the scenario cannot be ruled out a priori, so the analysis in this section naturally complements the discussion about bias.

It is worth stressing that negatively biased consultants are not "unbiased liars" (discussed briefly in Section 5), i.e. they do not consistently produce falsidical judgments. Rather, a negatively biased consultant's strategy depends on researcher's expectation, just like in the case of positive bias. To achieve their goals it may not be enough for a negatively biased consultant to deliver a falsidical judgment, for it is possible that the falsidical judgment corroborates linguist's expectations and this results in consultant's negative payoff. In principle, a more effective strategy is to guess the expected judgment of the linguist and deliver the opposite judgment. If the consultant has incomplete information about linguist's expectations, this situation results in **the negative consultant's dilemma** (NCD) summarized in Table 5. Similarly to Table 3 outlining the positive consultant dilemma, the first column features linguist's expectation about whether a linguistic expression has a property P, while the first row features judgments that the negatively biased consultant may deliver. From the point of view of the consultant, the linguist wins when the consultant delivers the judgment satisfying linguist's expectation, but unlike in the PCD this is precisely then when the consultant loses. In this case, the Nash equilibrium cannot be read off the table by means of the convenient rule of thumb used previously, since Table 5 has are no cells in which the first value is the column's maximum and the second value is the row's maximum. Nonetheless, the game has a mixed strategy Nash equilibrium, which can be calculated once the dilemma is transformed into a Bayesian game.

Table 5. Negative consultant's dilemma

	P	**¬P**
P	1, -1	-1, 1
¬P	-1, 1	1, -1

Given that the consultant does not know which report is expected by the linguist (or more technically: they do not know linguist's type), the best guess is distribute probability of type equally. Thus, reporting P with 50% probability that the linguists expects P yields 0 payoff (wins and losses cancel each other out).

$$-0.5 + (1 - 0.5) = 0$$

Analogically reporting ¬P with 50% probability that the linguist expects P yields 0 payoff.

$$0.5 – (1 – 0.5) = 0$$

Since neither strategy yields better payoffs, the negatively biased consultant's optimal strategy is to randomize with equal probability, i.e. to report P and ¬P in 50% of cases each. This supports the intuitive conclusion that the best strategy to deceive the linguist, provided the consultant does not know the linguist's expectations, is to deliver judgments at random.

It is worth repeating that there is no a priori reason to believe that a negatively biased consultant produces predominantly falsidical judgments. Consultant displaying either type of bias deliver judgments on the basis of their beliefs about researcher's expectations and the decision is logically independent from veridicality of the judgment. It may well be the case the biased judgments coincides with a veridical judgment and the statistical and game-theoretical analysis show that this is the case in 50% of cases. What is the best strategy for the linguist faced with a negatively biased consultant? Effectively, the researcher faces the linguist's dilemma (LD) summarized in Table 4, when maximum payoffs are accepting a veridical judgment and rejecting a falsidical judgments. Accepting or rejecting all judgments consistently yields positive payoffs in 50% in of cases each.

The conclusion from the analysis so far is that the optimal strategy is to accept all judgment of the unbiased consultant, 50% of judgment from the positively biased one, and 50% of judgments from the negatively biased one. The problem is, however, that consultants will not identify themselves as biased and may not even be aware that they harbor some sort of bias. How can a linguist know which strategy to adopt to maximize the yield of veridical judgments? As one may guess, this dilemma is also amenable to decision-theoretical analysis. Table 6 shows the payoff matrix combining all strategies discussed so far. Since the situation is characterized by incomplete information (the linguist does not know anything about consultant's putative bias), the Laplace criterion may be used to distribute probability of each type of consultant evenly (at 33% each). The strategy of accepting all judgments yields 100% of veridical judgments from unbiased consultants, 50% of veridical judgments from positively and negatively biased consultants each. Since in the biased cases the strategy also yields 50% of falsidical judgment, wins and losses cancel each other out and therefore acceptance yields 0 payoff in the cases with bias. Unbiased consultants deliver veridical judgments in all cases, so the linguist always receives maximum (1) payoff when accepting an unbiased judgment and suffers maximum loss when they reject it (-1).

Table 6. Payoff matrix for the elicitation decision making

	unbiased	**positively biased**	**negatively biased**
accept	**1**	0	0
reject	-1	0	0

Clearly, consistent acceptance is the best strategy, since it maximizes overall expected payoff:

$$0.33 + 0 + 0 = 0.33$$

compared to the expected overall payoff of consistent rejection:

$$-0.33 + 0 + 0 = -0.33$$

Table 6 also allows for estimating the overall average probability of linguist's success in the game given equal ratios of unbiased and biased consultants. The probability is calculated with the following formula, where:

- *A(v)* stands for the probability (on the 0-1 scale) of accepting a veridical judgment,
- *u* stands for the number of unbiased consultants in the elicitation procedure,
- *p* stands for the number of positively biased consultants in the elicitation procedure, and
- *n* stands for the number negative biased consultants in the elicitation procedure.

$$A(v) = \frac{u + \frac{(p+n)}{2}}{u + p + n}$$

Under equal distribution of each type of consultants,*A(v)* equals 0.67.

4. Should we trust ourselves?

Linguists who do not object to elicitation of intuitions as a matter of principle tend to accept judgments from "naive" consultants, but many will draw the line at judgments elicited by the linguist herself. They argue that linguists are not "naive" enough to deliver reliable intuitions, since not only do they possess expert knowledge that may distort the intuitions, but also (perhaps more importantly) have a personal stake in the result of the elicitation. As I understand, the objection does not necessarily seek to undermine the professional integrity of researchers by suggesting that they will deliberately distort their intuitions in favor of the desired outcome; it merely points out that linguists, despite their best efforts, may not be able to divest themselves of unconscious bias. At first glance, it may seem that the analysis in Section

3 lends support to this view. After all, in the scenarios with biased consultants the incomplete information about the expectations of the linguist forced the consultants to resort to mixed strategies and, as a consequence, deliver veridical judgments in as much as 50% of cases. I will argue, however, that in the case of linguist acting in good faith the putative unconscious biases do not change the structure of the game in any radical way and acceptance of judgments is still the optimal strategy.[4]

What is the main concern about bias during self-elicitation? As I understand it, linguists do not worry that expert knowledge in linguistics unavoidably cripples one's ability to evaluate linguistic expressions. Perhaps a linguist is no better than a naive consultant at evaluating, for example, well-formedness of an expression, but I see no reason to believe that a linguist is significantly worse in it only by the virtue of possessing expert knowledge. What is more worrying is that having a personal stake in the results of judgment elicitation, like the desire to have our predictions corroborated, unavoidably distorts our intuitions. It is, therefore, fruitful to think about the situation in terms of the question: "Would the linguist's judgment be the same if they had no personal stake in the outcome of the elicitation procedure?" This is the starting point for the analyses presented in this section and Two possible answers to this question define the states of affairs represented in the first row of the decision matrices below. The first column represents to possible decisions available to the linguist: t accept or to reject the judgment, just like in Section 3. Once again, I will consider three scenarios: lack of bias, positive bias, and negative bias.

4.1. Unbiased judgment

One possible objection at this juncture is that a linguist is always biased to a certain extent, so the first scenario should not be taken into account. Yet even if a linguist is always biased, I do not believe that the bias always has an important role in shaping a judgment. Consider the following sentence:

(2) Cat dog the the chased.

Consider also a linguist whose theoretical claim would be supported by evaluating (1) as ungrammatical and (2) as grammatical. Is it likely that unconscious positive bias would make the linguist genuinely believe that (1) is ungrammatical and (2) is grammatical? Perhaps this possibility cannot be ruled out a priori, but it does not appear very plausible. In clear cases of well-formedness and other linguistic properties unconscious bias is unlikely to seriously distort judgment. Even if we assume no professional integrity on the part of the linguist, they have to take into consideration the risk that a healthy community of researchers will recognize such fraudulent judgments, prevent publication of the results, and tarnish the linguist's reputation. It is clear even without preparing a payoff matrix that a conscious and deliberate lie of this sort is the losing strategy for the linguist. This shows that at least in some clear cases linguists are able to deliver honest and unbiased judgments and they have an incentive to do so, even when they cannot always divest themselves of bias entirely. Thus, the possibility of an unbiased judgment, i.e. a judgment undistorted by linguist's expectations, has some

irreducible probability and should be left in the equation.

Table 7 shows the payoff matrix for such an unbiased self-elicitation. Since in this kind of elicitation linguist's judgment is the same as naive consultant's judgment by definition, the column representing the possibility of an altered intuition is missing from the table (in other words, there is no such possibility in this scenario). Unsurprisingly, the decision that maximizes payoff is to accept the unbiased judgment in all cases.

Table 7. Payoff matrix for unbiased judgment

	Would the judgment differ?
	no
accept	**1**
reject	-1

Obviously, linguist's best choice in such a situation is to accept the judgment.

4.2. Positively biased judgments

The scenario with a positively biased linguist fuels most objections against self-elicitation. Simply speaking, positively biased linguists tend to produce judgments that support their expected outcome and the possibility that the biased judgment would differ outside elicitation procedure is a real threat. Note, however, that positive bias does guarantee that the judgment would change; there is also a possibility that the biased judgment would remain unaltered. The effect is analogical to the one discussed in Section 3.2: bias may stabilize an intuition rather than distort it. In such a case, rejecting the intuition is a patently wrong decisions and, in decision-theoretical terms, it minimizes linguist's payoff (see Table 8). The scenario is characterized by incomplete information: linguists do not know whether their judgment are stabilized or what the probability of a stabilized judgment is. Under the Laplace criterion, they may assume that the probabilities are distributed equally (50% for stabilized and 50% for altered judgments).

Table 8. Positively biased self-elicitation

	Would the judgment differ?	
	no	**yes**
accept	1	-1
reject	-1	1

While rejecting the altered judgment is certainly a win for the linguist, rejecting a stabilized judgment is certainly a loss, since in such a case the linguist has valuable piece of data at hand, yet they discard it. Similarly, accepting an altered judgment is linguist's loss, but accepting a stabilized intuition is a win. Since under both strategies ("reject" and "accept") wins and losses are equally likely, the expected payoff converges onto 0 in the long run.

4.3. Negatively biased judgments

Judging by the literature on the methodology of linguistics, the possibility of a negative biased linguist is not particularly worrying. For if negatively biased linguists exist, they are strange people indeed: they tend to produce judgments that defeat their own expectations. Yet even if we admit that there is a serious possibility that a negative bias alters judgments which a linguist would produce outside elicitation procedure, it cannot be assumed a priori that the bias guarantees a change of judgment in all cases. For example, if a negatively biased researcher expects an expression to have property P, they will be inclined to accept that the expression does not have P. Yet perhaps this is exactly what their judgment would be outside the elicitation procedure and therefore the bias merely stabilizes the veridical judgment. The payoff matrix for negatively biased elicitation is shown in Table 9 and is exactly the same as the matrix in Table 8 for positively biased self-elicitation. Obviously, with both types of bias the best strategy is to accept the judgment that would not change in a neutral context and to reject the one that would change. Since there is (once again) incomplete information about probability of each possibility, the best thing to do is to assume equal probability of both states of affairs (50% each). Consequently, the probability of accepting a veridical judgment is 50%, and the overall payoff converges onto 0 (due to 50% probability of accepting a falsidical judgment).

Table 9. Negatively biased self-elicitation

	Would the judgment differ?	
	no	**yes**
accept	1	-1
reject	-1	1

Just like in eliciting judgment from a consultant, in self-elicitation linguists do not know whether their judgments are biased or not and have to make a decision under uncertainty. The optimal strategies in the three above-mentioned scenarios are summarized into a payoff matrix in Table 10 and by applying the Laplace criterion it is assumed that all of the states of affairs are equally plausible.[5] Table 10 is identical with Table 6, which demonstrates that the final dilemma of the linguist ("to accept or to reject") in elicitation with a consultant is essentially the same as the dilemma in self-elicitation, at least in decision-theoretical terms. The optimal decision for the unbiased judgment is to consistently accept judgments, which

yields positive payoff in all cases of lack of bias and positive payoff in 50% of biased cases (offset by negative payoffs from accepting falsidical judgments). Expected overall payoff from consistent acceptance is once again 0.33, expected overall payoff from consistent rejection is -0.33, and the acceptance is the best strategy in 67% of cases.

Table 10. Payoff matrix in self-elicitation decision making

	unbiased	positively biased	negatively biased
accept	1	0	0
reject	-1	0	0

5. Psychological plausibility and other strategic adjustments

The mathematical model used in the above analyses is necessarily heavily idealized and may only approximate real-life situations. The formal model can be tinkered with in various ways to produce different results, but not all of the modification make the model a better approximation of reality. While in general game and decision theory steer clear of speculations about the psychology of decision-makers, two assumptions pertaining to psychological plausibility were made throughout Sections 3 and 4. The first in the already mentioned plausibility of an unbiased judgment during self-elicitation, especially about clear-case expressions. Another assumption is leaving out the “unbiased liar” scenario out of the equation.

An “unbiased liar” is a hypothetical consultant delivering uniformly falsidical judgments. An unbiased liar is not simply negatively biased, since a negatively biased consultant attempts to actively defeat researchers expectations rather than to deliver non-veridical judgments. If a negatively biased consultant reaches the conclusion that their veridical judgment defeats linguist’s expectations, the consultant will deliver a veridical judgment. In contrast, the judgment of an unbiased liar does not depend in any way on the liar’s beliefs about linguist’s expectations: their goal is to report a falsidical judgment at all times. From the mathematical point of view, the unbiased liar is a polar opposite of the unbiased consultant. If included in the model and ascribed probability equal to the probabilities of other types of consultants (under the Laplace criterion), the unbiased liar would offset the effect of the unbiased consultant and reduce the overall optimality of acceptance strategy to 50%. From the psychological point of view, however, a linguist is highly unlikely to work with an unbiased liar in real life. This is not to say that people never lie to each other, but when they do, they try to achieve some goal other than just saying something false. Typically, liars believe that lies will benefit them in one way or another. If a person has a choice between lying and telling the truth and the person has nothing to gain from either action, the person will probably tell the truth, if only because moral rules incentivize truthfulness or because the person is discouraged from lying by the fear of the lie being discovered. This again is not to say that unbiased liars do not exist and some people suffering from compulsive lying disorder come close to this hypothetical type. However, I do not believe that in real life the

unbiased liar scenario is so probable that including this option in the equation genuinely increases the the mathematical model's faithfulness to real-life situations. Not all possibilities are highly probable and while unbiased liars cannot be ruled out a priori, the same is true about the possibility that some consultants deliver judgment at random, so their choices are more adequately modeled as coin-tossing rather than strategic deliberations. Our experiences with people people (at least my experiences) do not lend much support for this claim that a significant number of people are unbiased liars and random decision-makers. An overwhelming majority of people are rational in the game-theoretical sense, i.e. they have certain goals and select strategies which help them achieve these goals; they do not select strategies at random or lie when the expected payoff from lying is relatively low.

The model in Sections 3 and 4 can be further adjusted with respect to psychological plausibility, so that it approximates real life more closely. One idealization used throughout the analysis is the Laplace criterion. When there was no way of knowing the probability of the unbiased consultant scenario relative to the two biased consultants scenarios, it was assumed that probabilities were distributed equally among the scenarios. This assumption seems to go against psychological plausibility by ascribing too much relative probability to the negative bias. If probabilities are indeed equal, one out of three consultants would actively plot against or at least semi-consciously attempt to defeat linguist's expectations. While I am ready to accept that this is precisely what some consultants do, it seems unlikely that they are as numerous as unbiased consultants, who simply "say what they think," and overly compliant positive biased consultants. Provided that consultants have no personal stake in a particular outcome of elicitation procedure, it is hard to see why one out of three persons should try to sabotage linguist's expectations. The point is even more striking when self-elicitation is considered. Here, the main factor threatening the veridicality of judgment is probably confirmation bias resulting in positive rather than negative prejudice. While it is easy to believe in what one wishes to be the case, it is much harder to believe in what we do not wish to be the case. Thus, a linguist may be inclined to produce intuitive judgments favorable to their theoretical claims, but why should we assume that they are equally inclined to produce judgments that refute their theoretical claims? I do not wish to argue that negative bias never takes place, but it appears that the mathematical model would be more realistic if the probability of the negative bias were demoted relative to positive bias and lack of bias. Decreasing the probability of the negative bias results in the overall increase of the optimality of acceptance as linguist's preferred strategy. For example, assuming that the ratio of unbiased to positively biased to negatively biased consultants is 2 to 2 to 1 (and hence the probability of negatively biased consultant is 20% rather than 33%), optimality of acceptance increases to 70%.

Another idealization is the assumption that the payoff from accepting a veridical judgment is equal to the payoff from rejecting a falsidical judgment; throughout the article the value of both the payoff is 1. This does not appear to be the case in real life. When linguists accept veridical judgments, not only have they obtained valuable data, but they can also use that data to develop their theories. When linguists merely reject falsidical judgments, they have avoided accepting erroneous data, but they remain without any data useful in their research. Let us assume that only 50% of intuitions are veridical. If linguist whose research relies crucially on consultants' judgments chooses to accept all judgments, as much as 50% of the data will be unreliable, but they will also make progress in their research and may hope to eradicate the adverse effect of unreliable data at a later stage. If a linguist chooses to reject all judgments, they will successfully defend

themselves against all falsidical intuitions, but their research will not even get off the ground due to lack of data. Thus, the payoffs from consistent acceptance appear to be greater than the payoffs from consistent rejection and the cost of rejecting veridical judgments outweighs the cost of accepting falsidical ones. If payoffs for acceptance and rejection are adjusted in the linguist's dilemma payoff matrix (dictating the optimal strategy for handling biased judgments under uncertainty; Table 11), acceptance becomes the best strategy: even though the rejection of altered judgments is still a Nash equilibrium, it is an inefficient equilibrium which yields lower payoffs than the equilibrium of accepting the stabilized judgment.

Table 11. Linguist's dilemma adjusted for profits

	stabilized	**altered**
accept	**1**	-1
reject	-1	**0.9**

Yet another possible complaint is that the analysis pays too little attention to the fact that most consultants' biases are unconscious and cannot be modeled as deliberate strategic choices, as in Section 3. This objection is correct, but not fatal. In real life, neither positively, nor negatively biased consultants are cunning conspirators calculating probabilities of various strategies to maximize their payoffs, although this idealization makes mathematical analysis easier. Notice, however, that decision theory used throughout Section 3 and 4 to analyze the linguist's dilemma construes judgments as states rather than outcomes of strategic planning. Obviously, consultants' judgments are not liveless states either: they are delivered by active agents in dynamic situations. Yet the idealization offered by the game theory (consultants as deliberate strategists) and decision theory (consultants as static states of affairs) are two extremes of a spectrum embracing biases of all shades of self-awareness and intentionality. Notice also, that both frameworks offer the same conclusions reached in very similar ways. Thus, the uncertainty about the degree of awareness and deliberation of bias is merely the uncertainty about which formal framework approximates the bias more closely, but it does not undermine the overall conclusion that consistent acceptance is researcher's optimal strategy during elicitation.

Finally, it may be argued that the mathematical model construes various decisions throughout elicitation in binary terms, while in real life the decisions are more gradable. Linguists may simply accept or reject a judgment, but they may also accept it with some reservations or, more generally, trust and distrust the consultant to various extent. Consultants, in turn, may deliver judgments and comment that they are not entirely certain whether the expression under investigation has the property P or not (e.g. whether a sentence is grammatical or ungrammatical). After all, this is the very reason why linguists use prefixes "??" and "*" to distinguish dubious expressions from straightforwardly ungrammatical ones. From the mathematical point of view, there are ways to accommodate gradability into the game-theoretical model. One of them is to link the payoff values to the degree of linguist's trust and consultant's confidence; for example, accepting a less confident

judgment with a lower degree of trust could result in lower linguist's payoff (e.g. 70) than accepting a highly confident judgment with a higher degree of trust (e.g. 90). In order to implement this solution, the consultant could be asked to evaluate their degree confidence of their judgment (for instance) on a 1-5 scale and the linguist could mark their trust about the consultant's judgment on a similar scale. The researcher would also need to design a mathematical function assigning payoff values to the combination of linguist's trust and consultant's confidence.

Would the introduction of gradability change the overall conclusion that consistent acceptance is linguist's best strategy? Much depends on the way gradability is implemented in the mathematical model. However, it seems unlikely that replacing binary judgments and acceptance with gradable alternatives without significantly changing other elements of the model would significantly reduce the optimality of acceptance. Acceptance is optimal, because it is always a good decision to trust unbiased consultants and there is no difference between trusting or distrusting the biased consultants (both strategies yield 0 payoffs). Whether confidence is gradable or not is logically independent of this mechanism.

6. Conclusion

Accepting consultants' judgments is the optimal strategy in elicitation with "naive" consultants and in self-elicitation in 67% of cases, assuming that unbiased, positively biases, and negatively biased consultants are equally likely. Is that much? That depends. On the one hand, indiscriminate acceptance is not dramatically better than acceptance at random; for example, when the linguist chooses to accept or reject by tossing a coin (a veridical judgment would be accepted in 50% of cases). This is certainly far from impressive. On the other hand, however, the danger of putative biases inherent in elicitation is much smaller than many linguists tend to believe. If the above model is correct and it adequately approximates real-life situations, it is hard to argue that the consultants should be distrusted as a matter of principle, even during self-elicitation. On the contrary, they provide veridical judgments in a clear majority of cases.

It should be also borne in mind that 67% is the result obtained before psychological plausibility and strategic factors were taken into account. Most probably, the group of negatively biased consultants should "weigh" less statistically relative to the two other groups, especially in self-elicitation. Moreover, in the grand scheme of things, acceptance is more rewarding than rejection, even if some falsidical judgments are accepted in the process, since rejection yields no veridical judgments at all. Both of the adjustments increase the optimality of consistent acceptance as a strategy in elicitation procedures.

Footnotes

[1] The ability to produce judgments is a part of Ryle's (2009 [1949]) implicit "knowledge-how" as opposed explicit "knowledge-that." See Polanyi (1962, 2009) for the related notion "tacit knowledge".

[2] One author who appears to hold such a view is Ewa Dąbrowska: "the use of the analyst's own introspective judgments is problematic for two reasons. First, many aspects of our mental life are not accessible to introspection, and these,

obviously, have to be studied using different methods" (Dąbrowska 2016, 480). The other reason signaled in the passage is bias: the proper of topic of this article.

[3] Technically, the consultant resorts to the so-called a maximin strategy, attempting to maximize one player's payoff assuming that the opponent is attempting to minimize the player's payoff.

[4] One caveat is that the argument does not apply to the situation when a dishonest researcher deliberately delivers falsidical intuitive judgments to find support for certain theoretical claims. My argument applies only to linguists acting in good faith who may nonetheless fall prey to unconscious bias.

[5] Readers convinced that some kind of bias is always present during self-elicitation should bear in mind that the probability of unbiased self-elicitation cannot be eliminated a priori. As the discussion on the sentences (1) and (2) demonstrates, bias is unlikely to distort judgment about clear-case expressions. At best, one could propose that the probability of unbiased judgment should be reduced in the cases of controversial expressions, but this cannot be proposed as a universal solution without any consideration for whether the expression used in research is indeed controversial.

References

- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge: The MIT Press.
- Clark, Robin. 2011. *Meaningful Games: Exploring Language with Game Theory*. Cambridge: MIT Press.
- Coseriu, Eugenio. 1991. *El hombre y su lenguaje*. Madrid: Editorial Gredos.
- Dąbrowska, Ewa. 2016. "Cognitive Linguistics' Seven Deadly Sins."*Cognitive Linguistics* 27 (4): 479–91. https://doi.org/10.1515/cog-2016-0059
- Danziger, Kurt. 1980. "The History of Introspection Reconsidered."*Journal of the History of the Behavioral Sciences*16 (3): 241–62.
- Gibbs, Jr., Raymond W. 2006. "Introspection and Cognitive Linguistics: Should We Trust Our Own Intuitions?"*Annual Review of Cognitive Linguistics*. https://doi.org/10.1075/arcl.4.06gib
- Guala, Francesco. 2016. "Philosophy of the Social Sciences: Naturalism and Antinaturalism in the Philosophy of Social Science." In *The Oxford Handbook of Philosophy of Science*, edited by Paul Humphreys, 34–64. New York: Oxford University Press.
- Halliday, Michael A. K., Angus McIntosh, and Peter Strevens. 1964.*Linguistic Sciences and Language Teaching*. London: Longman.
- Halliday, Michael, and Christian Matthiessen. 2004. *An Introduction to Functional Grammar*. London : New York: Routledge.
- Itkonen, Esa. 2005. "Concerning the Synthesis between Intuition-Based Study of Norms and Observation-Based Study of Corpora." *SKY Journal of Linguistics* 18: 357–77.
- Itkonen, Isa. 2008. "Concerning the Role of Consciousness in Linguistics."*Journal of Consciousness Studies*15 (6): 15–33.

- Jaeger, Gerhard. 2008. “Applications of Game Theory in Linguistics.”*Language and Linguistics Compass* 2 (3): 406–21. https://doi.org/10.1111/j.1749-818X.2008.00053.x
- Langacker, Ronald W. 2008. *Cognitive Grammar. A Basic Introduction*. New York: Oxford University Press.
- Pietarinen, Ahti-Veikko, ed. 2007. *Game Theory and Linguistic Meaning*. Amsterdam: Elsevier.
- Polanyi, Michael. 1962. *Personal Knowledge: Towards a Post-Critical Philosophy*. Chicago: University of Chicago Press.
- ———. 2009. *The Tacit Dimension*. Chicago-London: University of Chicago Press.
- Ryle, Gilbert. 2009 [1949]. *The Concept of Mind: 60th Anniversary Edition*. New York: Routledge.
- Saussure, Ferdinand de. 1966 [1916]. *Course in General Linguistics*. Translated by Wade Baskin. New York, Toronto and London: McGraw-Hill.
- Schütze, Carson T. 2016. *The Empirical Base of Linguistics Grammaticality Judgments and Linguistic Methodology*. Berlin: Language Science Press.
- Wasow, Thomas, and Jennifer Arnold. 2005. “Intuitions in Linguistic Argumentation.”*Lingua* 115 (11): 1481–96. https://doi.org/10.1016/j.lingua.2004.07.001