Studies in Second Language Acquisition, **42** (2020), 919–932
doi:10.1017/S0272263120000182

Research Report

EXPLORING THE VERIDICALITY AND REACTIVITY OF SUBJECTIVE MEASURES OF AWARENESS

IS A "GUESS" REALLY A GUESS?

Rebecca Sachs *

Virginia International University

Phillip Hamrick

Kent State University

Timothy J. McCormick

Georgetown University

Ronald P. Leow

Georgetown University

Abstract

Subjective measures (SMs) of awareness assume (a) participants can accurately report the implicit/explicit status of their knowledge and (b) the act of reporting does not change that knowledge. However, SMs suffer from nonveridicality (e.g., overreporting of "guess" responses) and reactivity (e.g., prompting rule search). Attempting to improve the validity of "guess" responses, we conducted an exploratory mixed-methods replication of Rebuschat et al. (2013). Participants (N = 30) were randomly assigned to Traditional, True Guess, and NoSMs conditions. True Guess participants were led to believe the computer would replace "guess" responses with random answers. Confirming that SMs are reactive, Traditional and True Guess participants responded more slowly and accurately, with greater awareness of the linguistic target. Moreover, although True Guess participants responded "guess" less frequently, interviews revealed this was due not to greater veridicality, but rather to additional reactivity. We conclude with directions for further research to enhance the validity of SMs.

The authors would like to thank Stephanie Leow, Nymisha Mattapalli, and Van To for their assistance in data collection and transcription, and the reviewers for their very helpful comments on a previous version of this manuscript.

* Correspondence concerning this article should be addressed to Rebecca Sachs, School of Education, Virginia International University, 4401 Village Drive, Fairfax, Virginia 22030. Email: rsachs@viu.edu

INTRODUCTION

In second language acquisition (SLA), as in cognitive psychology, questions surrounding implicit learning have long been debated from theoretical (e.g., Leow, 2015a; Schmidt, 1990 and elsewhere; Tomlin & Villa, 1994) and empirical perspectives (e.g., Hama & Leow, 2010; Leow, 2000; Williams, 2005). If no solid consensus on the role of (un)awareness in language learning has emerged, this may be due in part to the difficulties involved in operationalizing and measuring the construct (Leow et al., 2011). A review of approaches to operationalizing awareness reveals two broad stages that can be conceptualized as representing the *process* and the *product* of learning, respectively. The first stage (process) involves receiving, processing, encoding, and/or accessing information online, in real time, and is typically measured using concurrent think-aloud protocols (e.g., Leow, 2000; Hama & Leow, 2010), which have been critiqued for their potential reactivity. The second stage (product) involves the nonconcurrent retrieval of stored linguistic knowledge and is typically measured using retrospective interviews or questionnaires (e.g., Leung & Williams, 2011; Williams, 2005). The major critiques of such off-line measures often involve veridicality, given concerns about memory decay, fabrication, reluctance to report low-confidence knowledge, or an inability to differentiate or capture the lowest levels of awareness (see Leow, 2015b).

The internal validity problems entailed in the use of off-line measures (Leow & Hama, 2013) have featured prominently in a debate at the heart of the study of implicit learning in SLA: whether novel form-meaning mappings can be learned without awareness. In a seminal study in this area, Williams (2005) embedded four novel artificial determiners encoding distance and animacy in English sentences. While distance was trained overtly, animacy served as a hidden regularity. Finding that some participants were able to select the appropriate determiner (out of two options) after training without being able to characterize the animacy pattern accurately in retrospective interviews, Williams concluded that learning without awareness might be possible for adult L2 learners. Some subsequent conceptual replications corroborated this finding (e.g., Leung & Williams, 2011; 2012), but others failed to do so (e.g., Faretta-Stutenberg & Morgan-Short, 2011; Hama & Leow, 2010), and still others found evidence of both implicit and explicit knowledge (e.g., Rebuschat et al., 2013; Rebuschat et al., 2015) by employing both online and off-line verbal reports along with subjective measures of awareness, imported from cognitive science (e.g., Dienes & Scott, 2005).

SUBJECTIVE MEASURES OF AWARENESS

A stated benefit of subjective measures (SMs) of awareness is that they allow researchers to identify implicit knowledge even in the presence of explicit knowledge (Rebuschat, 2013; see, e.g., Hamrick & Rebuschat, 2012; 2014; Rebuschat et al., 2013; 2015). SMs include confidence ratings and source attributions. In confidence ratings, participants indicate their level of confidence in each test response, often on a Likert scale including options such as *not at all confident*, *somewhat confident*, *very confident*, and *absolutely certain*. In source attributions, participants identify whether each test response is based

https://doi.org/10.1017/S0272263120000182 Published online by Cambridge University Press

on a *guess*, *intuition*, *memory*, or a *rule*. Participants are instructed to respond "guess" only if they might as well have flipped a coin with a 50-50 chance of being correct. If their accuracy is significantly above chance on no-confidence or guess responses, the participants are said to possess implicit judgment knowledge or implicit structural knowledge, respectively. In other words, they are considered to be unaware *that* they know and/or unaware of *what* they know, and the inference is that they presumably learned without awareness (Dienes & Scott, 2005). Two assumptions underlying the use of SMs are (a) that participants can accurately report the nature of their knowledge and (b) that the act of doing so does not affect the nature of that knowledge. However, patterns of results observed in studies employing SMs raise major validity concerns regarding these assumptions.

THE PRESENT STUDY

In this short research report, which represents the first step in a larger-scale project currently underway, we describe our initial attempt to address two high-priority concerns that have arisen in our own work (e.g., Rebuschat et al., 2013; 2015): specifically, the nonveridicality of guess attributions and the reactivity of source attributions. Regarding veridicality, comparisons of participants' source attributions and confidence ratings have revealed that participants sometimes respond "guess" on the same items where they indicate some degree of confidence—an inherent contradiction that implies nonveridicality in the source attributions, confidence ratings, or both. In postexperiment interviews, participants sometimes admit that they have used the "guess" option to represent low-confidence hunches or vague recollections that did not feel strong or clear enough to be attributed to intuition or memory. Additionally, participants sometimes report understanding the concept of guessing to include educated guesses that involve some degree of conscious analysis, reflecting how the term is used in common parlance. In none of these cases should above-chance performance on guess attributions be considered valid evidence of implicit knowledge. As for reactivity, concerns arise from the fact that asking participants to think metacognitively about their responses can sometimes stimulate awareness or change the nature of their knowledge. For example, use of the word "rule" in source attributions has been found to prompt participants to engage in active rule-search behavior during the test (Rebuschat et al., 2015).

Although validity concerns arise with other aspects of subjective measures as well, we decided to focus first on "guess" responses. At the extreme end of the implicit/explicit continuum, they might be considered to provide the most unambiguous evidence of implicit knowledge, yet the reality seems to be that they represent a grab bag of low-confidence intuitions, memories, and rules as well. As will become clear, this preliminary exploration revealed challenges beyond those we had previously encountered and pointed toward future modifications that might be more successful in improving the internal validity of research in this area. We used both quantitative and qualitative methods to answer the following research questions:

1. Veridicality: Can we reduce the rate of nonveridical guess attributions by modifying the instructions for the source attributions?
2. Reactivity: Are subjective measures reactive? If so, what are some of the sources of reactivity?

https://doi.org/10.1017/S0272263120000182 Published online by Cambridge University Press

METHODS

PARTICIPANTS

Thirty university students (15 women, $M_{age} = 22.44$, range: 18–35) were randomly assigned into three groups: Traditional SMs (i.e., the same as those used in Rebuschat et al., 2013; 2015; $n = 10$), True Guess (i.e., with SMs modified to give the impression that guesses would be replaced with random computer-generated responses, $n = 10$), and NoSMs (without any subjective measures, $n = 10$). In other words, the key manipulations that differentiated these groups were (a) the presence or absence of SMs and (b) the type of language used for the no-confidence and guess response options. Additional information on the participants in each group, including their language backgrounds and number of linguistics courses they had taken, is available in Table S1 of the Supplementary Materials online.

MATERIALS AND PROCEDURE

Semi-artificial language learning experiment

Following previous replications, we employed the artificial determiner system from Williams (2005), in which four artificial determiners (*gi*, *ro*, *ul*, *ne*) encode both distance (near vs. far) and animacy (animate vs. inanimate). The determiners *gi* and *ro* refer to near entities that are animate and inanimate, respectively. Likewise, *ul* and *ne* refer to far entities that are animate and inanimate, respectively. Apart from the aforementioned differences in the use of SMs in the nontraditional conditions and the addition of some interview questions relevant to the focus of the present study, the materials and procedures were exactly the same as those used in the non-think-aloud experimental conditions of Rebuschat et al. (2013; 2015). The training and test sentences are available in the IRIS digital repository (https://www.iris-database.org/iris/app/home/detail?id=york%3a807980; Marsden et al., 2016).

All participants were pretrained on the near-far meanings of the novel determiners without any mention of animacy. Then, in the exposure phase, they were instructed to read a series of 144 sentences aloud. For each sentence, they had to decide whether the determiner referred to a near or far entity, then repeat the novel determiner and its noun (e.g., "*gi* bears") aloud while forming a mental image of the situation. After this exposure phase, participants were given a two-alternative forced-choice task as an unannounced test. In the test phase, after four practice items, they read 36 new sentences, each containing a blank in place of the artificial determiner, and had to choose which of two determiners (e.g., *gi* or *ro*) seemed "more familiar, better, or more appropriate based on what you have done so far." In these test sentences, there were equal numbers of trained, partially trained, and generalization items for each determiner. Critically, the test options were always matched for distance, so participants could only get the answer correct by choosing the option with the correct animacy value.

Subjective measures of awareness

SMs (confidence ratings and source attributions) were administered on each test trial for the Traditional and True Guess groups, but not for the NoSMs group. Whereas the instructions for the Traditional group were identical to those used in Rebuschat et al.

https://doi.org/10.1017/S0272263120000182 Published online by Cambridge University Press

(2013; 2015), True Guess participants received modified instructions for no-confidence and guess responses. In the Traditional condition, the definition of guessing referred to the flipping of a coin to convey the idea of a random response. However, Rebuschat et al. had found this insufficient to ensure that participants would respond in accordance with the intended meaning; for instance, participants sometimes attributed their answers to guesses alongside indications of at least some confidence. Thus, to lead participants to select "guess" only when their responses truly did feel random, the True Guess instructions—in a ruse—led participants to believe that if they selected either "no confidence" or "guess," the computer would randomly select an answer for them instead. To clarify and emphasize this, the instructions explained that if a random response was just as likely to be correct, participants should "be fine with the computer generating the answer for you … with a 50-50 chance of being correct (the same as if you had truly guessed)." Then, during the test, whenever True Guess participants indicated zero confidence or a guess, they were informed that the computer had randomly selected an answer for them, followed by a prompt to indicate whether or not they accepted this. The exact wording of these instructions is available in the Supplementary Materials online.

The idea behind the ruse was that, if participants had any reason to prefer one answer over another (which would not count as a guess in researchers' intended sense), then they would be less likely to relinquish control to the computer. Accordingly, allowing the computer to guess would indicate a true guess. Through this manipulation, we sought to reduce the number of educated guesses and low-confidence non-guesses that participants would classify (nonveridically) as "guess" responses. In so doing, we hoped to develop a technique that could at least partially increase the validity of researchers' claims to have found evidence of implicit knowledge.

Postexperimental interview

Immediately after the test phase, participants were interviewed individually to probe how they had reacted to the test phase, whether and how often they had guessed, how they defined guessing, whether their guess attributions had matched their definitions, how they had felt about guessing and whether they had ever avoided it for any reason, what criteria they had used to make their choices, whether they had ever responded based on a rule and if so what its content was, and whether (and if so when) they had become aware of the hidden animacy regularity. The full set of interview questions is available in the Supplementary Materials online. Those related to awareness followed the procedures described in Rebuschat et al. (2013; 2015). Two interview audio recordings were lost due to equipment issues.[1]

To begin, two researchers independently coded nine of the interviews (three per experimental condition) with a focus on 10 variables derived from the questions outlined in the preceding text. There was 100% agreement on the coding of seven variables and moderate to very good agreement ($0.60 < \kappa_w < 0.90$) on the other three. For the variables with complete agreement, the researchers divided the rest of the data to code separately. For the three with less agreement, following a debriefing, both researchers coded all the interviews and then discussed any discrepancies to decide the final codes. This article reports on the accuracy of participants' definitions of guessing ($\kappa_w = 0.90$), their feelings

https://doi.org/10.1017/S0272263120000182 Published online by Cambridge University Press

about and avoidance of guessing (100% agreement), their reported attempts to formulate rules ($\kappa_w = 0.60$), and their awareness of animacy (100% agreement).

RESULTS

VERIDICALITY

To examine whether we had succeeded in reducing the proportion of test items attributed to guesses by modifying the instructions for the SMs, we compared the average proportions of zero-confidence and guess attributions the participants made in the Traditional and True Guess groups.[2] The descriptive statistics including outliers are available in Tables S2 and S3 of the Supplementary Materials. For the inferential analyses presented here, one Traditional and two True Guess participants were removed for being greater than 2.0 standard deviations (SD) from their group means. A 2×2 mixed ANOVA with SMs (2 levels: confidence rating, source attribution) as a within-subjects factor and Group (2 levels: Traditional, True Guess) as a between-subjects factor revealed a statistically significant effect of SM type, $F(1, 15) = 9.94$, $p = .007$, $\eta^2 = 0.39$, with a slightly lower proportion of zero-confidence ratings ($M = 8.76\%$) than guess attributions ($M = 9.07\%$) overall; a significant effect of Group, $F(1, 15) = 9.94$, $p = .032$, $\eta^2 = .27$, with a lower proportion of zero-confidence and guess attributions in the True Guess group ($M = 7.75\%$ and 2.33%, respectively) than in the Traditional group ($M = 8.58\%$ and 15.04%, respectively); and no significant interaction, $F(1, 15) = 2.88$, $p = .11$. Notably, in the True Guess group (minus the outliers), there were only seven guesses in total, made by six participants. These results reveal two important points: first, the effect of SM type, apparently driven by the pattern of results in the Traditional group, indicates that a greater proportion of responses were attributed to "guess" than to "zero confidence." That is, Traditional participants expressed at least some degree of confidence for some of the responses they attributed to guesses. Because the definition of guessing assumed by subjective measures entails zero confidence, this suggests that the validity of at least some guess responses in the Traditional group might be called into question. Second, these results indicate that the True Guess participants were significantly less likely to attribute test responses to guessing than the Traditional participants were, with the true mean effect (95% CI) between 3.12% and 21.39% reduction in the guess response rate.

To understand why True Guess participants attributed a lower proportion of their responses to guessing, we inspected the patterns of responses to the interview questions regarding participants' feelings about guessing and avoidance of guessing.[3] As Table 1 indicates, seven participants in the Traditional condition reported having felt fine about guessing, as represented by P(articipant)18's statement that "I really didn't know, so I was kind of like, OK, just choose one" and P47's laughing response that he felt "very nonchalant, I guess." In contrast, seven participants in the True Guess condition reported having felt bad about guessing or even having avoided it. Excerpts from P41's comments are especially telling in this regard:

Um (laughs) I felt kinda bad, like I hadn't really figured out what I was, uh, supposed to be doing yet…. It said we put in, um, the computer response, is that fine? And I was like, no, that's not fine

https://doi.org/10.1017/S0272263120000182 Published online by Cambridge University Press

TABLE 1. Participants' reported feelings about guessing according to experimental condition

	Traditional ($n = 9$)	True guess ($n = 10$)	No subjective measures ($n = 9$)
Avoided guessing	1	3	0
Felt bad about guessing or reluctant to guess	1	4	3
Felt fine about guessing	7	3	6

Note: $p = .15$ (Fisher's exact test).

(laughs), and so I picked no. So then I just did my own guess as opposed to letting the computer pick one for me…. I think after that I sort of was like, OK, let me think about, um, some sort of way I can strategically go about this and *not* be guessing. So, um, answering guess just sort of put me in a position to be like, OK, let me figure out some sort of rule I can follow or put a system there so I don't have to answer guess (laughs).

Reports of avoiding guessing were not restricted to True Guess participants. In the Traditional condition, P15 admitted, "I didn't want to use too many guesses," and P20 explained that she had avoided responding guess "because my mother didn't raise a punk; I like to give it a try." However, these were the only two Traditional participants who expressed any qualms about making use of the "guess" option.

In sum, by modifying the instructions for the subjective measures, we seem to have been able to reduce the rate of potentially nonveridical guess attributions in the True Guess condition; however, we may also have reduced the rate of valid guess responses by making many of the True Guess participants feel reluctant to select that option independently of the perceived status of their knowledge. In effect, not only was our True Guess manipulation apparently too heavy-handed to solve the veridicality problem but it also created a new reactivity problem. Worse, as shown in Table 2, it did this without leading to more accurate definitions of guessing among True Guess participants.

In assessing the accuracy of participants' definitions of guessing, we considered their interview responses to be on target if they referred to flipping a coin (corresponding to the Traditional instructions), letting the computer answer (corresponding to the True Guess instructions), or having no idea and/or answering randomly or blindly. We classified off-target responses according to the types of mischaracterization they represented. These included referring to intuition, memory, or knowledge; an inability to understand the rule or pattern (without providing an accurate definition of guessing); or a combination of these.

In the NoSMs condition, without the benefit of instructions regarding guessing, only three participants responded with definitions that would have matched implicit-learning researchers' intentions. Four made references to using a gut feeling, intuition, or instinct, or picking what sounded right, compared to only one participant each holding that misconception in the Traditional and True Guess conditions. Moreover, two NoSMs participants referred to making an educated guess, as illustrated by these comments from P44:

Well, I guess there are different levels of guessing. There is guessing when you have no clue, it could be either and you really don't know, so it really is more of an eeny meeny miney mo, and then the

https://doi.org/10.1017/S0272263120000182 Published online by Cambridge University Press

TABLE 2. Accuracy of participants' definitions of guessing according to experimental condition

	Traditional ($n = 9$)	True guess ($n = 9$)	No subjective measures ($n = 10$)
On target (total)	6	5	3
Flipping a coin	1	1	0
Letting the computer answer	0	0	0
Having no idea, answering randomly or blindly	5	4	3
Off target (total)	3	4	7
Using a gut feeling, intuition, or instinct; picking what sounded right	1	1	4
Using memory or knowledge (e.g., to make an educated guess)	1	0	2
Not understanding the rule, pattern, or association	0	2	0
Off target in multiple ways	1	1	1

Note: $p = .32$ (Fisher's exact test).

further level of guessing, more of an educated guess, like I remember hearing "ul bees" before; therefore, I'm going to go with that one again because that's the closest thing I have.

Comparing the True Guess and Traditional conditions against each other, the accuracy of participants' definitions of guessing was fairly similar. Only one participant in each condition referred to flipping a coin, and no one made references to letting the computer answer. Instead, five or six participants in each SMs group described answering randomly, having no idea how to respond, or not having anything to go on, as in P39's description of taking "a shot in the dark" (True Guess) or P50's statement "it's kind of like just blindly choosing something" (Traditional). Three Traditional participants reported off-target definitions of guessing, with one referring to intuition, another referring to educated guesses, and another giving a definition that was off-target in multiple ways. Similarly, four True Guess participants had off-target definitions, with one referring to intuition, two describing an inability to create an association or understand the rule, and one making multiple off-target comments. In other words, our attempt to ensure that participants in the True Guess condition would internalize a more accurate definition of guessing does not seem to have been successful.

REACTIVITY

The veridicality results indicated that leading participants to believe that the computer would guess for them made participants not only less likely to report guessing, but also perhaps more likely to feel reluctant about guessing, in some cases reportedly prompting a search for rules to avoid having to guess. This in itself suggests a reactivity effect. To determine whether the SMs were reactive and explore some of the potential sources of reactivity, we conducted quantitative analyses of overall test accuracy and reaction times across groups as well as qualitative analyses of participants' reported attempts to formulate rules and their awareness of the hidden animacy regularity. For the statistical analyses, two outliers were removed for being greater than +/– 2.0 SD from their group

https://doi.org/10.1017/S0272263120000182 Published online by Cambridge University Press

accuracy means. A one-way ANOVA on mean accuracy by group revealed a statistically significant effect of Group, $F(2, 14.03) = 4.19$, $p = .037$, with both the True Guess group ($M = 66.70\%$, $SD = 21.70\%$), Welch's $t(10.76) = 2.14$, $p = .055$, 95% CI [0.42%, 31.22%], and the Traditional group ($M = 63.33\%$, $SD = 14.37\%$), $t(17) = 2.37$, $p = .029$, 95% CI [13.41%, 22.72%] showing significantly higher accuracy than the NoSMs group ($M = 53.30\%$, $SD = 6.80\%$). A one-way ANOVA on reaction times by group also revealed a statistically significant effect of Group, $F(2, 30) = 3.67$, $p = .03$, whereby the True Guess ($M = 6103$, $SD = 2245$), $t(17) = 10.80$, $p < .001$, and Traditional groups ($M = 6863$, $SD = 2577$), $t(17) = 12.72$, $p < .001$, were slower at responding to test trials than the NoSMs group ($M = 4456$, $SD = 1298$). Taken together, these findings indicate that employing SMs led to more accurate[4] and slower test performance relative to participants who were not asked to produce subjective measures of awareness.

The qualitative interview data shed further light on these patterns. Table 3 displays how many participants expressed various levels of awareness of the hidden animacy regularity across groups, while Table 4 shows how many participants reported attempts to formulate rules. To characterize the participants' levels of awareness, we employed Rebuschat et al.'s (2015) coding criteria (pp. 312–313). The interview data revealed complete awareness of the animacy regularity 2.5 times as frequently in the SMs groups compared to the NoSMs group. Moreover, while six or seven of the participants in each SMs condition displayed at least minimal awareness of animacy, six of the participants in the NoSMs condition showed no such evidence whatsoever.

TABLE 3. Awareness of the hidden animacy regularity according to experimental condition

	Traditional ($n = 10$)	True guess ($n = 9$)	No subjective measures ($n = 10$)
Complete awareness, expressed with confidence	3	3	1
Complete awareness, but expressed with hesitance	2	2	1
At least partial awareness	1	1	1
Minimal awareness	1	0	1
Lack of awareness	3	3	6

Note: $p = .87$ (Fisher's exact test).

TABLE 4. Participants' reported attempts to formulate rules according to experimental condition

	Traditional ($n = 9$)	True guess ($n = 9$)	No subjective measures ($n = 10$)
Proactively attempted to formulate a rule	5	3	0
Recognized a rule or somewhat attempted to formulate one	4	2	4
Did not attempt to formulate a rule	0	4	6

Note: $p = .01$ (Fisher's exact test).

https://doi.org/10.1017/S0272263120000182 Published online by Cambridge University Press

Likely due in part to the appearance of the word "rule" among the choices for the source attributions, participants in the SMs conditions frequently referred to the existence of a rule they assumed they were supposed to come up with. A common reaction to the test phase in those groups is illustrated by P36's comment, "I hadn't really been paying much attention to the rule so it kind of threw me off." Indeed, all the participants in the Traditional group reported either recognizing a rule or at least somewhat attempting to formulate one. Interestingly, in comparison to the matter-of-fact approach reported by many Traditional participants, the emotionality of True Guess participants' responses was often striking, as Table 1 on their feelings about guessing also suggested. For instance, whereas P47 (Traditional) straightforwardly recounted his process of coming up with a rule, P14 (True Guess) admitted that his drive to formulate a rule derived from feeling somewhat frustrated or cornered by the test phase:

P47 (Traditional): I felt like y'all was like trying to test, uh, memory at first, and then I realized that, uh, there was more of a rule, and I was trying to figure out y'all rule, but then I had to make up my own to do the question. [Researcher asks when] Um … probably at the end, uh, when they asked us to fill it in because you had to make the rule.
P14 (True Guess): Um, I was guessing, it was at the start of the final section, and I felt slightly frustrated or I felt kind of cornered, and I had to come up with a rule very quickly.

In contrast, representative quotations from the NoSMs group suggested that NoSMs participants "felt like it was more of a guessing game in the second part" (P22) and "just went with the one that seemed best" (P34) or "just [went] by memory … [or] like, instinct, I guess" (P43). In fact, P25 reported feeling "kind of relieved when I didn't have to say anything" (after reading sentences aloud during the exposure phase) and "just tried to put in the word that sounded right." Instead of referring to a rule, NoSMs participants were more likely to make comments along the lines of P28's admission that "I just, like, guessed on all of them because, yeah, I didn't know I had to memorize, like, which ones go in which places." Furthermore, among the participants classified as unaware, those in the NoSMs condition tended to report having based their test responses on what "sounded better" or "felt more familiar," whereas SMs participants who had remained unaware of the animacy regularity were more likely to refer to alternative rules they had fabricated. For instance, all three unaware Traditional participants and one unaware True Guess participant mentioned a singular/plural rule, while the other unaware True Guess participants reported attempts to use mnemonics or hypotheses they had generated regarding sentence structure or positive versus negative situations.

One puzzling aspect of the results relates to the fact that, despite similar patterns of awareness and similar accuracy rates in the SMs conditions for the various types of source attributions (presented in Table S3 of the Supplementary Materials), the proportions of responses that the Traditional and True Guess participants attributed to rules differed by roughly 21%. Even though all the Traditional participants reported rule recognition and/or rule search in the interviews, they attributed their responses to rules only 15% of the time in the test phase, indicating three times as often that their responses were based on intuition (46%), while attributing 23% of responses to memory and 15% to guesses. In the True Guess condition, however, despite having reported less rule search in the interviews, participants attributed their test responses to rules 36% of the time, along with fairly

https://doi.org/10.1017/S0272263120000182 Published online by Cambridge University Press

similar proportions of intuition (28%) and memory (30%) responses, and with only 3% of responses attributed to guesses. These discrepancies raise further questions about veridicality.

DISCUSSION

As a first step in our endeavor to improve the validity of subjective measures of awareness, we attempted to (a) enhance the veridicality of "guess" responses by manipulating the instructions for confidence ratings and source attributions, and (b) gain a deeper understanding of sources of reactivity by quantitatively and qualitatively comparing groups of participants whose test phase either did or did not include SMs. More specifically, we sought to reduce the number of nonveridical guess attributions by leading True Guess participants to believe that the answers they attributed to guesses would be replaced by randomly generated computer decisions. This method may have overshot its target. Although the proportion of guess attributions was substantially lower (as predicted) in the True Guess condition compared to the Traditional condition, interview data suggested that this was not due to more accurate definitions of guessing among True Guess participants, but rather due to a reluctance among 70% of them to have a computer guess for them, compared to the 78% of Traditional participants who expressed that they felt fine about guessing. Moreover, our True Guess manipulation seems to have created a new reactivity problem in the sense that some True Guess participants reported having engaged in rule search precisely because they did not want to guess. Beyond the possible reactivity of the test phase (see Hamrick & Sachs, 2018), additional sources of reactivity were found in the inclusion of subjective measures, which led participants to respond more slowly and with greater accuracy. This is likely related to the fact that, compared to the NoSMs group, SMs participants were more likely to attempt to formulate rules (Traditional 100%, True Guess 56%, NoSMs 40%) and to develop awareness (Traditional 70%, True Guess 67%, NoSMs 40%). In part, as also reported by Rebuschat et al. (2015), this may have been due to the presence of the term "rule" in the instructions.

These results have important implications for the study of (un)awareness in SLA and underscore the need to continue to critique and improve the internal validity of studies that use subjective measures. Although SMs hold out the potential benefit of tapping implicit and explicit knowledge separately on different test trials, we must contend with the veridicality and reactivity concerns they exhibit, namely, the possibility that participants' "guess" attributions may not be valid, and the possibility that presenting participants with the idea that a rule may exist might lead some of them to search proactively for a pattern. SLA research employing SMs has been interpreted as providing evidence of both implicit and explicit knowledge (e.g., Hamrick & Rebuschat, 2014; Rebuschat & Williams, 2012; Rebuschat et al., 2013), but the present findings cast doubt on the basis for those interpretations. There is an irony here: SMs are often touted as a sensitive tool for investigating implicit knowledge, but their reactivity may make participants more likely to develop explicit knowledge.

This study also demonstrates the value of triangulating measures of awareness (cf., e.g., Rebuschat et al., 2015). SMs, online verbal reports (think-alouds), and off-line verbal reports (interviews, questionnaires) involve different advantages, challenges, and

https://doi.org/10.1017/S0272263120000182 Published online by Cambridge University Press

limitations, and can produce different results. For example, in this experiment, groups of participants who displayed different rates of guess and rule responses during the test phase nonetheless showed the same levels of awareness according to the interviews. Not only can triangulation help to enhance the internal validity of research, but it can also illuminate possible reasons for patterns of results and uncover discrepancies that future studies can be designed to explain. Based on the quantitative and qualitative data gathered in this preliminary study, we have begun to explore the effects of (a) operationalizing the "guess" option in a more familiar and intuitive way that involves flipping a coin, and (b) removing the "rule" option and replacing it with a less suggestive alternative (i.e., "other"). We hope that these new experimental conditions will produce more veridical data with less reactivity.

LIMITATIONS, CONCLUSIONS, AND FUTURE DIRECTIONS

In addition to the usual cautions against generalizing warranted by considerations of sampling and statistical power, which we are currently addressing in an expanded study with more conditions and more participants, there are reasons to avoid applying our attempted validity improvement. Not least of these is the fact that, in seeking to minimize nonveridical guess attributions, we introduced a novel source of reactivity that led some participants to be averse to guessing. As part of our larger-scale investigation, we are not only continuing to gather qualitative data through interviews but also measuring individual differences (IDs) in locus of control (LOC) and attitudes toward computers as possible mediators of participants' willingness to guess or to hand over their prerogative to choose. In brief, LOC involves people's beliefs about whether events in their lives tend to be caused by external factors (e.g., fate, luck, chance) or internal factors (e.g., personal effort, ability) and their own actions (Rotter, 1966; see Halpert & Hill, 2011). If LOC or attitudes toward computers correlate with participants' frequency of attributing test responses to guesses, that may argue for including such ID measures as covariates when using source attributions to establish evidence of implicit knowledge.

All in all, the results of this replication and extension have confirmed that, due to the potential for both nonveridicality and reactivity, research that employs subjective measures to investigate implicit and explicit learning and knowledge should be interpreted with appropriate caution. While much remains to be done to enhance the validity of subjective measures of awareness, the findings of this preliminary study have demonstrated the value of taking incremental steps with mixed methods to reveal the unexpected consequences of experimental manipulations before launching a full-scale investigation. Even small-scale replications such as the one reported here can be quite informative and useful in prioritizing and guiding the allocation of resources for larger projects.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0272263120000182.

NOTES

[1]In a small number of cases, it was possible to code confidently for a variable based on written notes the interviewer had taken. This accounts for the varying group sizes in the tables.

[2]Although the purpose of this study was not to investigate the development of implicit knowledge, the Supplementary Materials include traditional analyses of the subjective measures, with Figure S1 displaying overall accuracy in each group and Tables S2 and S3 displaying the proportions and accuracy of responses with different source attributions and confidence ratings. With so few instances of guessing in the True Guess condition and no subjective measures in the NoSMs condition, it was not possible to examine the accuracy of guess responses as a way of investigating whether there was evidence of implicit knowledge in those conditions; however, the data show that the participants in this experiment performed comparably to what has been reported in other studies, where participants in conditions similar to our Traditional condition had mean accuracy scores between 52% to 76%. The mean score and interquartile range for our Traditional group fall within that range. It is important to note that outliers were not excluded from the data presented in the Supplementary Materials.

[3]As an efficient way of summarizing our qualitative data for a short report, we have included tables displaying how we coded participants' interview responses to give readers a sense of group trends. Although we were not testing specific null hypotheses in these qualitative analyses, we have provided the results of Fisher's exact tests in response to a reviewer's request.

[4]We thank a reviewer who suggested that we examine whether learning during the test phase differed across our three groups. Space does not permit a full reporting of these results, but indeed, using multilevel modeling, we found a statistically significant positive relationship between test trial number and accuracy in both the True Guess and Traditional groups, but not in the NoSMs group. This suggests that True Guess and Traditional participants learned to a modest degree during testing, while the NoSMs group did not, providing further evidence of the reactivity of subjective measures.

REFERENCES

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research*, *69*, 338–351.

Faretta-Stutenberg, M., & Morgan-Short, K. (2011). Learning without awareness reconsidered: A replication of Williams (2005). In G. Granena, J. Koeth, S. Lee-Ellis, A. Lukyanchenko, G. P. Botana, & E. Rhoades (Eds.), *Selected proceedings of the 2010 Second Language Research Forum* (pp. 18–28). Cascadilla Proceedings Project.

Halpert, R., & Hill, R. (2011). *28 measures of locus of control.* https://www.yumpu.com/en/document/read/11010080/28-measures-of-locus-of-control-the-webs-most-useful-information

Hama, M., & Leow, R. P. (2010). Learning without awareness revisited: Extending Williams (2005). *Studies in Second Language Acquisition*, *32*, 465–491.

Hamrick, P., & Rebuschat, P. (2012). How implicit is statistical learning? In P. Rebuschat & J. Williams (Ed.), *Statistical learning and language acquisition* (pp. 365–382). De Gruyter Mouton.

Hamrick, P., & Rebuschat, P. (2014). Frequency effects, learning conditions, and the development of implicit and explicit lexical knowledge. In J. Connor-Linton & L. Amoroso (Eds.) *Measured language: Quantitative approaches to acquisition, assessment, processing, and variation* (pp. 125–140). Georgetown University Press.

Hamrick, P., & Sachs, R. (2018). Establishing evidence of learning in experiments employing artificial linguistic systems. *Studies in Second Language Acquisition*, *40*, 153–169.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior: Aware versus unaware learners. *Studies in Second Language Acquisition*, *22*, 557–584.

Leow, R. P. (2015a). *Explicit learning in the L2 classroom: A student-centered approach*. Routledge.

Leow, R. P. (2015b). Implicit learning in SLA: Of processes and products. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 47–65). John Benjamins.

Leow, R. P., & Hama, M. (2013). Implicit learning in SLA and the issue of internal validity: A response to Leung and Williams' "The implicit learning of mappings between forms and contextually derived meanings". *Studies in Second Language Acquisition*, *35*, 545–557.

Leow, R. P., Johnson, E., & Zárate-Sández, G. (2011). Getting a grip on the slippery construct of awareness: Toward a finer-grained methodological perspective. In C. Sanz & R. P. Leow (Eds.), *Implicit and explicit conditions, processes and knowledge in SLA and bilingualism* (pp. 61–72). Georgetown University Press.

Leung, J. H. C., & Williams, J. N. (2011). The implicit learning of mappings between forms and contextually derived meanings. *Studies in Second Language Acquisition*, *33*, 33–55.

Leung, J. H. C., & Williams, J. N. (2012). Constraints on implicit learning of grammatical form-meaning connections. *Language Learning*, *62*, 634–662.

Marsden, E., Mackey A., & Plonsky, L. (2016). The IRIS repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 1–21). Routledge. https://www.iris-database.org

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research: A review. *Language Learning*, *63*, 595–626.

Rebuschat, P., & Williams, J. (2012). Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics*, *33*, 829–856.

Rebuschat, P., Hamrick, P., Riestenberg, K., Sachs, R., & Ziegler, N. (2015). Triangulating measures of awareness: A contribution to the debate on learning without awareness. *Studies in Second Language Acquisition*, *37*, 299–334.

Rebuschat, P., Hamrick, P., Sachs, R., Riestenberg, K., & Ziegler, N. (2013) Implicit and explicit knowledge of form meaning connections: Evidence from subjective measures of awareness. In J. Bergsleithner, S. Frota, & J. K. Yoshioka (Eds.), *Noticing: L2 studies and essays in honor of Dick Schmidt* (pp. 255–275). University of Hawaii Press.

Rotter, J. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, *80*, 609.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 129–158.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition*, *38*, 293–316.

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition*, *27*, 269–304.