Assumptions of a subjective measure of consciousness: Three mappings

Zoltán Dienes

Department of Psychology

University of Sussex

Josef Perner

Department of Psychology

University of Salzburg

To appear in Rocco Gennaro (Ed.) "Higher order theories of consciousness"

John Benjamins Publishers.

Zoltán Dienes, Department of Psychology, School of Life Sciences, Sussex University, Brighton, BN1 9QG, UK.
E-mail: dienes@biols.susx.ac.uk
Tel: 00 44 273 678550
Home page: http://www.biols.susx.ac.uk/home/Zoltan_Dienes/

Josef Perner, Department of Psychology, University of Salzburg, Hellbrunnerstrasse 34, A-5020 Salzburg, Austria
email: josef.perner@sbg.ac.at
Tel. 0043-(0)662-8044-5124
Home page: http://www.sbg.ac.at/psy/people/perner_e.htm

Running head: SUBJECTIVE MEASURES OF CONSCIOUSNESS

Consider some event in the world, for example there being an object moving up. There are two distinct ways of knowing this event: being merely aware of it, or being consciously aware of it. We are aware of the event when the event influences cognitive processing or behaviour, we are sensitive to the event in some way. But just because we are sensitive to the event, that does not mean we are consciously aware of it. Similarly, we might know some regularity in the world, e.g. we may be behaviourally sensitive to the perceptual cues that indicate whether a chick is male or female. But that does not mean we consciously know the regularity. Experimental psychologists have debated for over a century how to determine whether we can be aware of stimuli or regularities (sensitive to them in some way) without being consciously aware of them (e.g. Peirce and Jastrow, 1884). The debate has been heated (e.g. Holender, 1986; Shanks & St John, 1994; Dienes & Perner, 1999; Merikle & Daneman, 1998), and we believe higher order theories can help provide clarity in evaluating useful criteria for deciding when people consciously know rather than simply know. Rosenthal (e.g. 1986, 2000, forthcoming, this volume) suggests we adopt a common intuition that we consciously know (that something is so) when we are conscious of the mental state by which we know. It is difficult to regard a person as consciously seeing if, despite 100% correct discrimination performance on a task, the person vigorously and adamantly insists that they see nothing, they are just guessing, they are not conscious of seeing at all (Weiskrantz, 1988, 1997). Thus, Weiskrantz argued that conscious seeing required a separate commentary by the person on the visual processing of the stimulus, i.e. on the mental state of seeing itself. Similarly, Rosenthal suggested that a mental state is conscious when we have a roughly contemporaneous thought (a higher order thought) to the effect that we are in

that state. Carruthers (1992, 2000, this volume) likewise argued that to understand phenomena such as blindsight we need to be able to distinguish worldly subjectivity (representing the world, a first order representation) and experiential subjectivity (representing one's mental states), the latter providing conscious awareness.

For adult humans, seeing is understood as a means of knowing; merely guessing that something is so cannot be a case of seeing that it is so. Thus, on higher order theory, when a mental state of e.g. seeing (that something is so) is a case of consciously seeing (that something is so), we (adults) would be able to distinguish between whether we merely guessed that something was the case, or that we knew it to some extent. Similarly, for any other occurrent knowledge state: To know consciously entails we can distinguish between knowing (to some degree) and guessing.

Higher order theories lead naturally to criteria for assessing conscious knowledge, namely criteria based on people's ability to determine the mental state that they are in, and that provides the knowledge (so called "subjective measures" of consciousness). Criteria based on merely the person's ability to determine the stimulus that was shown ("objective measures" of consciousness) could easily lead to false positives: First-order mental states allow the discrimination of stimuli (indeed, that is generally their function) and second order states are not needed at all for discrimination. Objective discrimination of stimuli entails that you are aware of the stimuli, but not that you are consciously aware of them. Subjective measures, which directly test for the existence of second order states, thus (according to the theory) directly test for the existence of conscious awareness.

One such subjective criterion is the "zero confidence-accuracy relationship criterion", normally called the "zero correlation criterion" for short (Dienes & Berry,

1997). When the subject makes a judgement, ask the subject to distinguish between guessing and different degrees of knowing. If the judgment expresses conscious knowledge - on those cases when it is knowledge and not guessing - then the subject should give a higher confidence rating when she actually knows the answer and a lower confidence rating when she is just guessing. In other words, conscious knowledge would prima facie be revealed by a correlation between confidence and accuracy, and unconscious knowledge by no correlation (the person does not know when she is guessing and when she is applying knowledge).

Another criterion based on a person's ability to determine the mental state that they are in is the "guessing criterion" (Dienes et al , 1995). Take all the cases where a person says they are guessing, and see if they are actually demonstrating the use of knowledge. This is the criterion that is satisfied in cases of blindsight (Weiskrantz, 1988, 1997). The person insists they are just guessing, but they can be discriminating up to 90-100% correct. So, based on the guessing criterion, the knowledge in blindsight is unconscious.

While the zero correlation and guessing criteria have face validity as measures of conscious awareness, are there conditions in which the criteria would give the wrong answer? What do we need to assume to ensure their validity? Dienes (submitted) used Rosenthal's higher order thought theory to flesh out some assumptions behind both the zero correlation and guessing criteria, regarding the extent to which the criteria could be biased[1]. In this paper, we will use higher order

[1] Essentially, bias in the way higher order thoughts are formed from first order states does not lead to subjective measures being biased measures of the conscious status of the first order states; but bias in measuring a second order thought does lead to biased measurement of the conscious status of first order states.

thought theory to consider some further assumptions required for using the zero correlation criterion[2].

We will first summarize the contexts in which the zero-correlation criterion has been previously applied, and then consider in detail its application to, first, implicit learning, and then subliminal perception.

Previous use of the zero correlation criterion

The zero-correlation criterion has now been applied extensively in the implicit learning literature. Implicit learning occurs when people learn by acquiring unconscious knowledge (for reviews see Shanks & St John, 1994; Dienes & Berry, 1997; Cleeremans, Destrebecqz, & Boyer, 1998). The term "implicit learning"was introduced by Reber (1967). Reber asked subjects to memorize strings of letters, where, unbeknownst to subjects, the order of letters within the string was constrained by a complex set of rules (i.e. an artificial grammar). After a few minutes of memorizing strings, the subjects were told about the existence of the rules (but not what they were) and asked to classify new strings as obeying the rules or not. Reber (see Reber, 1993, for a review of his work) found that subjects could classify new strings 60-70% correctly on average, while finding it difficult to say what the rules were that guided their performance. He argued the knowledge was unconscious. But, starting with Dulany, Carlson, and Dewey (1984), critics have been unhappy with free

[2] We will assume an actualist higher order theory, like Rosenthal's, in which the first order state and the higher order state are different representations. Thus, the Carruthers (1992, 2000, this volume) potentialist theory will be inconsistent with some of the arguments that follow, because it postulates the first order state does not require a separate second order representation for the first order state to be conscious. While we will refer to higher order thoughts, consistent with Rosenthal's theory, regarding the higher order states as thoughts rather than perceptions (e.g., Armstrong, 1980) is irrelevant for the arguments that follow.

report as an indicator of unconscious knowledge. Free report gives the subject the option of not stating some knowledge if they choose not to (by virtue of not being certain enough of it); and if the free report is requested some time after the decision, the subject might momentarily forget some of the bits of knowledge they brought to bear on the task.

Chan (1992) elicited a confidence rating in each classification decision, and showed subjects were no more confident in correct than incorrect decisions. Dienes et al (1995), Dienes and Altmann (1997), Allwood, Granhag, and Johansson (2000), Channon et al (2002), Tunney and Altmann (2001) and Dienes and Perner (1993) replicated these results, finding some conditions under which there was no within-subject relationship between confidence and accuracy. We argued this indicated subjects could not discriminate between mental states providing knowledge and e.g. those just corresponding to guessing; hence, the mental states were unconscious (see also Kelley, Burton, Kato, & Akamatsu, 2001, and Newell & Bright, 2002, who used the same lack of relationship between confidence and accuracy to argue for the use of unconscious knowledge in other learning paradigms). The method has an advantage over free report in that low confidence is no longer a means by which relevant conscious knowledge is excluded from measurement; rather the confidence itself becomes the object of study and can be directly assessed on every trial.

Kolb and Braun (1995) and Kunimoto, Miller, and Pashler (2001) applied a similar methodology to perception. Kolb and Braun investigated texture discrimination and Kunimoto et al word perception. They both found conditions under which confidence was not related to accuracy, and argued this demonstrated the existence of unconscious perception.

We will now consider some conditions that should be satisfied for the zero-correlation criterion to provide valid answers about the conscious status of mental states. After some preliminary considerations, we will consider implicit learning in detail, and then subliminal perception.

Preliminary considerations: attitudes and higher order thoughts

Human beings meet the world with a highly evolved learning system. Over evolutionary history this learning system has faced environments with specific statistical and other kinds of structures. When it meets a new environment or structured domain it can implicitly presume the structure will belong to a certain class of structure types and the need is to determine the parameters that specify that class of structure. In a first encounter with a domain recognized as novel in important respects, there will be uncertainty as to the right parameter values. If we are asked to decide whether a chick is male or female by looking at it, we may initially be just guessing. After much practice, we may make judgements with certainty. The system in making a judgement ("the chick is male"), has an attitude toward that judgement (guessing, knowing) that expresses itself in how the content is used, for example, the consistency with which it is used, the amount of contrary evidence that would overthrow it, etc. In between the attitudes of guessing and knowing there can be degrees of confidence, or thinking with some conviction. The English language does not express these attitudes very well; if one says there is an attitude of knowing or guessing, in everyday usage it may imply there must be second or even third order thoughts about guessing or knowing. But for the purposes of this paper consider

guessing, thinking with some conviction etc as being first order mental states, regardless of any higher order thoughts that may or may not be present. There can be an attitude (knowing, guessing, etc) independently of being aware of that attitude.

Consider, for example, one possible way in which an attitude may show itself as a particular attitude (independently of any awareness of the attitude), namely in the consistency with which it is used. If I am merely guessing that a given chick is male, on repeated presentations of the same chick (without knowing that it is the same chick) I may respond "male" as often as "female". The fact that I say "male" 50% of the time is an expression of my complete uncertainty as to whether the chick is a male. However, if I was certain it was male, I should say "male" 100% of the time. Similarly, an attitude in between guessing and certainty may express itself by my saying "male" e.g. 60% of the time. In this case, I would be engaging in a type of probability matching (like animals and people do in many but not all situations, Reber, 1989; Shanks, Tunney, & McCarthy, 2002) - matching relative frequency of response to a subjective probability. My tendency to say "male" 60% of the time would reflect the extent of my uncertainty for the judgement: I am somewhat sure that the chick is male, but I am not certain. In sum, there may be situations in which the consistency with which a judgment is made indicates how well the system has normally learnt that the judgment accurately represents the world.

Note that if we are using consistency to measure attitude, we cannot tell from any one trial what the attitude is. We need an ensemble of trials. On any one trial, there will be a certain attitude, but we won't know the attitude from just one trial. Over a set of trials we can estimate the attitude. In order to use consistency as a measure of attitude, we need to assume representational stability over the time span we are investigating. For example, one could be highly certain that Fluffy is male

one day (e.g. because we are told); and be highly certain that Fluffy is female the next day (because that is what we are told that day). In this case, certainty co-exists with inconsistency, but that's because the state of knowledge has changed from trial to trial. Only by presenting the object of the judgement repeatedly, while the person remains in the same state of knowledge, can consistency be used to measure attitude. We need to assume representational stability over the test phase.

To facilitate the arguments that follow, consider a state of affairs in the world, X. For example, X could be that "this chick is male". One forms a judgement (a mental state) about X with content *x*. For example, *x* could be "this chick is male". Call the attitude of the mental state ATT. So the full judgement can be written as ATT(*x*). For example, ATT(*x*) could be: I am fairly sure that "this chick is male", where ATT is fairly sure and *x* is "this chick is male". If the judgement is correct than *x* will correspond to X; if the judgement is incorrect, *x* will not correspond to X. This state of affairs is shown in Figure 1.

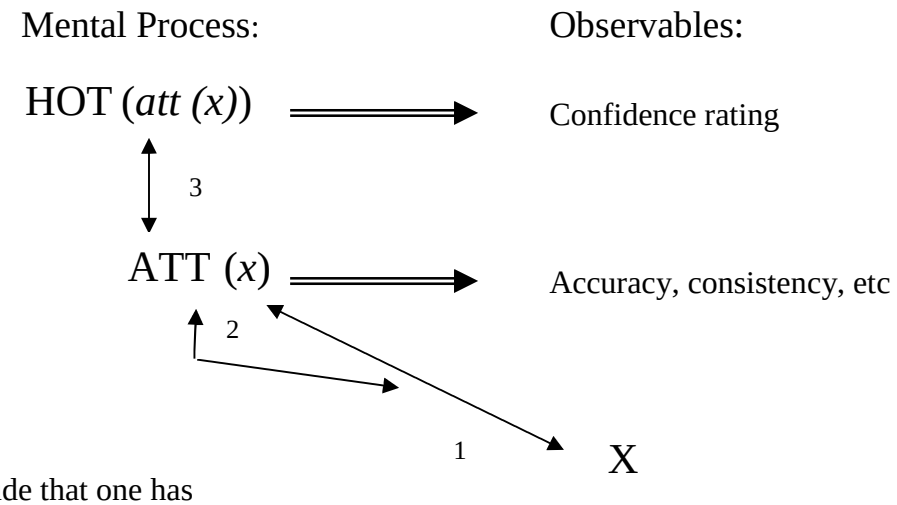


Figure 1. Relationship between the world, first order mental states, and higher order thoughts.

Consider a series of judgements, involving a number of different attitudes because I have different attitudes regarding the sex of different chicks; I may be certain about Donald, almost certain about Daffy, etc. Mapping (1) in the figure refers to the extent to which *x* corresponded to X over the series of test trials: What percentage of times over all judgments did I get it right? How often when I formed the judgement that a chick was a particular gender, was the chick that gender?

There is another relationship to consider (labelled 2 in the figure): Between the different attitudes and the percentage correct for each attitude. If the learning system is well adapted to the domain, then an attitude of certainty should be associated with a higher percentage of correct judgments than an attitude of lesser certainty would be. For example, if I am certain about Donald and only fairly sure about Daisy, am I correct more often for Donald than for Daisy?

Finally consider when the person has a higher order thought (HOT) which represents the person as having a certain attitude (*att*). For example, the HOT could represent that "I am guessing that this chick is male". One might represent oneself as guessing (*att*) even though one's actual attitude on that trial was one of being fairly sure (ATT). Mapping 3 refers to the extent to which the attitude, *att*, one represents oneself as having corresponds to the attitude ATT that one actually has.

When using the zero-correlation criterion, we wish to assess mapping 3: Are people aware of the attitude of knowing when they know something and guessing when they are just guessing? However, neither one's actual attitude (ATT) nor the higher order thought about one's attitude (*att*) can be strictly directly observed by the experimenter. Instead, as shown on the right hand side of Figure 1, the experimenter determines the subject's higher order thought by the subject's confidence ratings, and the subject's attitude by how correct the subject is. In the latter case, the assumption is

when the attitude involves more certainty, the subject will make more decisions correctly: the attitude of guessing should lead to chance performance and the attitude of knowing should lead to high performance. When is it valid to use these observables (confidence ratings and percent correct) to infer whether the quality of mapping 3 is better than zero? In other words, when can the zero-correlation criterion be used to infer the presence of conscious or unconscious mental states? That is the question we will consider in the rest of this paper.

Now we will use the mappings 1 to 3 illustrated in Figure 1 to consider the application of the zero-correlation criterion to implicit learning, and then to subliminal perception. In what follows we will need to consider a series of judgments. The zero-correlation criterion (and the guessing criterion) cannot be applied to a single judgment to determine if it was conscious or unconscious. It can only be applied to an ensemble of judgments to determine whether all of them were unconscious or whether at least some were conscious.

Applying the zero correlation criterion to implicit learning

Consider a subject learning an artificial grammar. After a training phase consisting of looking at grammatical strings, the subject is informed of the existence of a set of rules and asked to classify test strings. The subjects first order mental state for each judgment about a test string consists of two components: The content of the judgment (*x* in Figure 1, e.g. “this string is grammatical”) and the subject’s attitude to that content (ATT). Now we ask the subject for their second order thought (their confidence rating) which is their report (*att*) of their attitude used in the first order

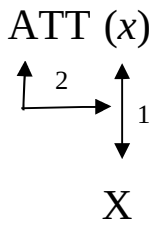
mental state. In assessing the conscious or unconscious status of the first order state, we wish to establish whether the content of the higher order thought (*att*) represents the actual attitude (ATT) used in the first order state (when subjects know, do they know that they are knowing?). But there are in fact three mappings to consider:

First, there is the mapping (1) of the content of the first order mental state onto the world (specifically, in this case, the world is the grammaticality of each test string as determined by the experimenter's grammar): $x \leftrightarrow X$ (mapping 1 in the figure). Researchers into the content of learning (e.g., in implicit learning paradigms, Tunney & Altmann, 1999; Cleeremans, 1993; Dienes & Fahey, 1995, 1998) are interested in investigating this mapping.

Consider a plot of the relationship between confidence and accuracy for a particular subject. Let us say it shows no relationship. Does this mean we have established the unconscious status of the subjects' knowledge? Not necessarily. Maybe the subject was mapping the attitude used in the first order state onto the content of their second order thoughts perfectly; but the content of the first order state was completely uncorrelated with the world. For example, the subject may have induced a grammar that would generate all training strings, but also half of the grammatical test strings and half of the non-grammatical test strings. On the basis of this grammar the subject would judge half the grammatical and half the non-grammatical test strings correctly the other half of each type of string incorrectly. As a result the subject would appear to be objectively guessing when in fact subjectively the subject was making 100% knowledgeable judgments. The experimenter had not come up with a psychologically plausible grammar; so the subject's highly evolved learning system presumed a different type of structure than the one the experimenter arbitrarily dreamt up. The learning system presupposed the "wrong" kind of model of

the world; but under normal conditions it would have learnt to some degree successfully at this juncture. So the subject has a first order attitude of “knowing” or “thinking with 100% conviction” on some trials, and this is reflected completely in the confidence rating. Nonetheless, there would be no relation between confidence and accuracy because the first order content-world mapping is not reliable. The first order state is a conscious mental state, but the zero-correlation criterion would indicate it is unconscious.

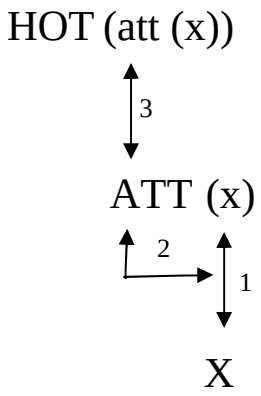
Second, there is the mapping (2) between the first order attitude (ATT), on the one hand, and the reliability of the first-order-content to world mapping (1), on the other:



Presumably, the first order attitude will normally map onto the reliability of the first-order-content to world mapping. Attitudes of greater certainty will generally be associated with contents that map the world reliably, when the learning system is adapted to the environment it is learning about. But of course this in no way guarantees that attitudes of certainty will always be towards contents that reliably map the world (even though they would under conditions to which the learning system is adapted). It does not even guarantee that in some restricted domain, variations in attitude correlate with variations in the reliability of content-world mapping. Researchers into metacognition are effectively interested in this mapping (e.g Reder & Ritter, 1992; Koriat, 1993, 1997; Gigerenzer, 2000): How does a person determine whether he or she has an appropriate attitude of knowing?

In an artificial grammar learning task a subject may have different attitudes towards the grammaticality of different test strings, and the subject may have accurate second order thoughts about those attitudes. But if the subject has as many percent correct choices for first order attitudes of high certainty as for attitudes of low certainty, there will be no correlation between confidence and accuracy despite appropriate higher order thoughts making all mental states conscious[3].

Third, there is the first order attitude – second order content mapping, which is the one we are actually interested in to assess the conscious or unconscious status of the first order states:



Mapping (3) is also of interest to people working in metacognition (e.g. Reder & Ritter, 1992; Koriat, 1993, 1997; Gigerenzer, 2000, as referenced for mapping (2)), who often are effectively interested in the joint effect of mappings (2) and (3), without distinguishing them

Our central question now is: How can we make sure that we are assessing mapping (3) with the zero correlation criterion, when mappings (1) and (2) also strongly affect the confidence-accuracy relationship?

[3] In terms of Gigerenzer's (e.g. 2000) theory, if the subject had induced cues and represented cue validities from the training set, but these cue validities were uncorrelated with the "ecological validities" of those same cues in the test set, then confidence would be unrelated to accuracy regardless of whether the subject was conscious of their first order attitude (as determined by the cue validities, on Gigerenzer's theory).

First, we have to make sure the subject has a minimal (non-zero) first order content-world mapping over the set of test items as a whole before applying the criterion. We can determine this simply by looking at subjects' percent correct classification. There would be no point looking at the strength of the confidence-accuracy relationship unless subjects have demonstrably induced knowledge that correlates with the experimenter's grammar; i.e. their percent correct classification is above baseline. This shows mapping (1) over all items as a whole to be above chance.

Next, in order to establish mapping (2), we need some way of measuring attitude. Let us presume that consistency of response directly reflects attitude. Since Reber (1967), people in the artificial grammar learning literature have often tested subjects twice on the same test strings to measure consistency: The proportion of strings classified twice correctly (CC), once correctly and once in error (CE), and twice in error (EE). Note that having a reliable mapping (2) requires variability in attitude (as observably indexed by e.g. response consistency). Typically, CC, EE, and CE are all non-negligible (e.g. CC is about 0.60, EE about .15, and CE about .25: Reber, 1989). This is important in showing variability in consistency. If subjects responded deterministically with a single rule that classified some items correctly and some incorrectly, being correct or incorrect to different items would not be related to different attitudes, because there may only be one attitude, i.e. of knowing (as pointed out by Dienes & Perner, 1996). In this case, CE would be zero and mapping (2) would be zero or undefined.

Given the minimal requirement of CE being non-negligible, how are we to establish whether there is a positive relationship between attitude and percent correct (i.e. mapping (2))? To recap, it appears subjects have different attitudes towards the judgments they make about the grammaticality of strings. For some strings, a subject

may respond “grammatical” and “non-grammatical” about equally often, reflecting the attitude of guessing. For other strings, subjects respond “grammatical” with some probability above 0.5, indicating a stronger attitude, one of greater certainty that the string is grammatical, or with a probability below 0.5, indicating greater certainty that the string is non-grammatical. How can we determine whether stronger attitudes are associated with greater accuracy?

Answering this question depends on how the zero-correlation criterion is to be measured. For example, let us say the zero-correlation criterion is to be measured by comparing the average percent correct for high confidence responses with average per cent correct for low confidence responses (as used by Tunney and Shanks, submitted). Consider four test items, and the subject is tested on each 100 times (with an ideal subject who never gets bored and has no memory of previous test trials). The subject gets item 1 correct 100 times, the subject gets item 2 correct 0 times, and he gets items 3 and 4 correct 80 times each. The overall percent correct is 65% (mapping 1 is fine). The subject has given 400 responses. If we take the 200 responses with the highest consistency (hence the stronger attitudes), this would be items 1 and 2, where the probability of emitting a given response is 1.0. The average percent correct for these two items is 50%. For the remaining 200 responses (i.e. to items 3 and 4), 80% of the responses are correct. Thus, average accuracy is higher for the weaker attitudes (80%) compared to the stronger attitudes (50%), as shown in Figure 2. Mapping (2) is negative. So even if subjects had completely accurate higher order thoughts (mapping 3), and hence completely conscious knowledge, the zero-correlation criterion would not show a positive relation between confidence and accuracy (as measured by: %correct given high confidence minus %correct given low confidence).

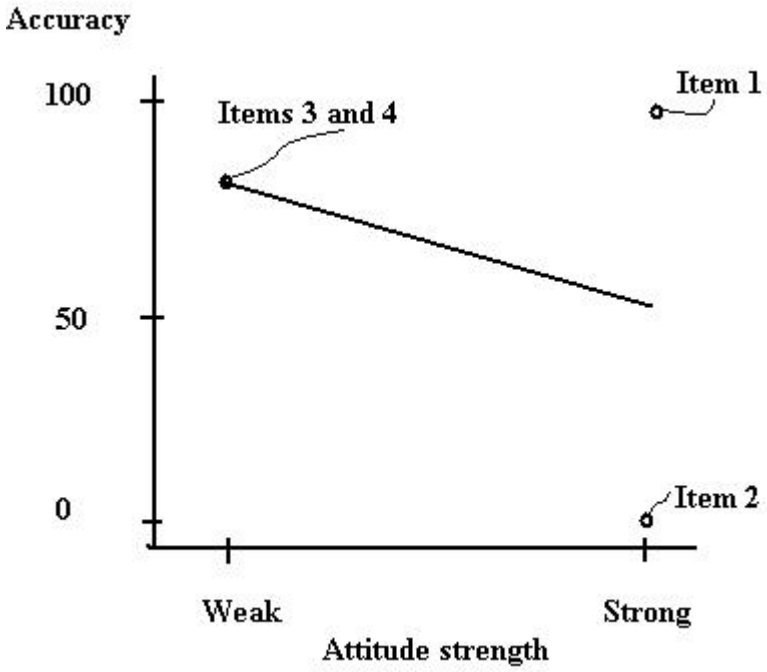


Figure 2

Example of mapping (2) being negative

We will now consider another measure of the quality of mapping (2). Chan (1992) measured the zero-correlation criterion by subtracting the average confidence for all incorrect responses from the average confidence for all correct responses (The "Chan difference score"). We could use the Chan difference score, but apply it to actual attitudes (ATT) rather than confidence ratings (as estimates of our HOTs (*att*) about our attitudes) in order to investigate mapping (2). That is, we subtract the average strength of attitude for correct responses from the average strength of attitude for incorrect responses. In the above example, the average attitude strength is 0.88 for correct responses and 0.77 for incorrect responses[4]. The difference is positive,

[4] The relevant calculations may be best understood after reading the Appendix. We assume that the proportion of times a stimulus elicits the correct response is the attitude towards the correct response; i.e. the attitude for the correct response is the proportion correct for that item. Conversely, the attitude

mapping (2) is satisfied. So if HOTs could directly represent attitude strength, the zero-correlation criterion would be positive (appropriately) by Chan's difference score applied to confidence ratings in this example[5].

The Appendix considers in detail the measured strength of mapping (2) assuming different response models and measures of mapping (2) (including the Chan difference score). It concludes that the strength of mapping (2) is normally positive as measured by the Chan difference score, and inversely related to EE (Reber's notation for being consistently in error). Mapping (2) needs to be carefully considered depending on the way the zero-correlation criterion is to be measured. In general, the lower the value of EE, the better mapping (2), at least for the Chan difference score (and - so we speculate - more generally).

Having established positive mappings (1) and (2), the zero-correlation criterion can be investigated. A zero correlation provides evidence for unconscious mental states; a positive correlation indicates the presence of at least some conscious mental states (without ruling out the possibility of the existence of at least some unconscious mental states) (see also Dienes & Perner, 2001, for one further proviso, namely that the confidence ratings should not arise from any inferences of which the subject is conscious).

---

towards the incorrect response is (1 – proportion correct). The calculations are: For correct responses, there are 100 responses of strength 1.0 from item 1, 80 of strength 0.8 from item 3 and another 80 of strength 0.8 from item 4. This weighted average is 0.88. For incorrect responses, there are 100 responses with strength 1.0 from item 2, 20 responses of strength 0.2 from item 3, and 20 responses of strength 0.2 from item 4, giving a weighted average of 0.77.

[5] The following may be best understood after reading the Appendix: In the example in the text of mapping (2) being negative, different attitudes were collapsed together into one category. Specifically, items 3 and 4 are correct 80% of the time, and so those correct judgments have an attitude strength of .80; they are incorrect 20% of the time, and so those incorrect judgments have an attitude strength of .20. In Figure 2, an attitude strength of .80 for the correct responses was lumped together with an attitude strength of .20 for the incorrect responses, both attitudes being put in the "weak attitude" category. Maybe collapsing over attitude strengths in this way was the problem with this measure of mapping (2). Maybe the Chan difference score showed mapping (2) to be positive because it respects all the distinctions between attitudes strengths, making full use of the data. However, for reasons that are unclear, Tunney and Shanks (submitted) found binary confidence ratings resulted in a more sensitive measure of the zero-correlation criterion than more fine-grained confidence ratings. As this is opposite to expected, we are running a study to explore the finding.

Under most experimental conditions, subjects are likely to have at least some conscious knowledge. In this case, the zero-correlation criterion can still be used to compare the relative amount of conscious and unconscious knowledge across two conditions (compare Jacoby's process dissociation framework). First, one would establish the two conditions had the same percent correct classification (mapping (1) is the same) and the same EE value (mapping (2) is the same). Under these conditions, differences in the confidence-accuracy relationship then indicate different mixes of conscious and unconscious knowledge across the two conditions. For example, Chan (1992) compared a group of subjects in training who just memorized strings with a group who searched for rules. At test they had the same (non-significantly different) percent correct classification - mapping (1) was the same. They also had the same average level of confidence. Reber (e.g. 1989) had previously shown that rule search compared to memorization instructions changes consistency by increasing EE; i.e. mapping (2) is worse for rule search than memorization subjects (see Appendix for justification of this logic)[6]. Nonetheless, Chan found that the rule search group had a stronger confidence-accuracy correlation than the memorization group (according to the Chan difference score amongst other measures); this must have been due to a better mapping (3) for rule search than memorization. Hence there is evidence for a greater use of unconscious knowledge in the memorization group than the rule search group.

Applying the zero correlation criterion to subliminal perception

[6] Another possibility is that subjects told to search for rules simply believed they should be extra consistent in responding to the same item, as compared to memorization subjects. However, in his 1989 review (p. 228), Reber found that the total amount of consistency (CC + EE) in rule search and memorization conditions is identical.

Now consider the application of these issues to subliminal perception. Kunimoto et al (2001) found conditions under which subjects were just as accurate in identifying stimuli for high confidence as low confidence decisions. They concluded that this indicated the presence of unconscious perceptual states (after carefully considering, and then rejecting, possible psychometric artifactual explanations of the dissociation between confidence and accuracy). What mapping assumptions does their conclusion rely on?

The task of the subject in Kunimoto et al (2001) experiments one and two was first to say what word was there and then to give a confidence rating. Perception presents a somewhat different situation from implicit learning. Arguably, in the first instance perception always involves the attitude of knowing towards whatever its content is. Even when we know what we see is an illusion, our vision still seems to present it to us as 100% fact. Nonetheless, the content of one's perception may not clearly be one of the contents the experimenter is forcing us to respond with (e.g. "the word is green" etc is allowed by the experimenter, but "A funny bunch of lines" is not a response option). So one considers the content of the perception as evidence for the different words, and converts that into an attitude towards an allowable content e.g. being 60% sure that the word is green. The logic of the zero-correlation criterion runs like this: (i) The more (conscious or unconscious) perceptual evidence one has, the greater one's accuracy; and (ii) the more conscious perceptual evidence one has, the greater one's confidence. So a positive relationship between confidence and accuracy indicates conscious perceptual evidence. Now let's consider the mapping assumptions, as shown in Figure 1, required for this logic to actually go through.

Mapping (1) is the same as in Figure 1: To what extent does the perceptual system produce contents that accurately represent the stimuli? If the perceptual system sometimes misrepresents one word as another, this content could produce confident inaccurate answers, reducing the confidence accuracy correlation, even when perception is quite conscious. Kunimoto et al (2001) established that their subjects classified above chance, so mapping (1) is not problematic for their experiments.

There is also a version of mapping (2) to consider. The perceptual representation used to judge what word was presented will have various conceptual and perceptual contents (for example, that "the word is red", or how bright the word is, etc), and other properties like fine-grainedness, time taken to form, or clarity, and so on. Mapping (2) is the extent to which those representational properties taken to indicate the extent to which the representation is accurate (vs misrepresenting) actually map onto to the extent to which the representation is accurate or misrepresenting.

We have probably evolved so that in normal conditions mapping (2) is very good; at least, we tend to think so: We usually feel we are pretty accurate in judging whether we have seen or failed to have seen something. But the system is not infallible; sometimes even in everyday life we have what seem to us to be clear percepts but in fact we have seen incorrectly. After all, the system cannot have a God's eye view of when it gets things right and when it gets things wrong. Mapping (2) might be compromised in two ways.

First, the first order discrimination and the assessment of representational accuracy may be based on at least some different representational properties. For example, a person may be conscious, somewhat dimly, of seeing the word green, and

this perception may have certain qualities the person is also aware of. The first order discrimination is made on the basis of the relevant content provided by the perceptual system, namely that the word is green. Now how is the person to provide a confidence rating? They may rely on vividness to guide attitude even when it is irrelevant to discriminating what word was presented. Then different confidence ratings can be produced for the same level of first order accuracy. Subjects presumably use properties for assessing accuracy that are correlated with accuracy under normal conditions. But so long as the correlation is not perfect, restricting the range of variability of the properties (as would be done in e.g. a subliminal perception experiment where accuracy is forced to vary over a limited range) could reduce the correlation to zero[7].

Second, consider the case where the first order discrimination and the assessment of representational accuracy are based on exactly the same representational properties (e.g. the mechanism considered by Kunimoto et al, 2001, Appendix B, "the ideal observer"; Björkman, Juslin, & Winman, 1993). Where there is a finite number of options (e.g. red vs green vs blue vs yellow), a simple decision procedure is to consider the activation of representations having each of the specified contents; whichever representation has the greatest activation is chosen, or is more likely to be chosen (first order discrimination) and the amount of differential activation can be used to determine the accuracy of the representation. Let's assume that mapping (1) is satisfied so that the first order discrimination is above chance. For mapping (2) to fail, the activation of e.g. the "red" representation must on average not be stronger when the word is "red" than when the red is not "red". Mapping (2) could

[7] Whittlesea, Brooks, and Westcott (1994), present a situation where a categorization decision was based on different information than the confidence in the same decision. Subjects could be biased to base one decision (e.g. categorization) on typicality of individual features and the other (e.g. confidence) on exemplar similarity, or vice versa.

fail in this way if there were inhibitory links between the representational choices (or, more generally, cleaning up processes) that force the system into four attractors: red, yellow, green, or blue. It will go into the right attractor more often than the wrong one (mapping (1) satisfied), but when it is in the wrong attractor, the final activation level of the chosen representation will on average be the same, regardless of whether it is right or wrong (mapping (2) fails). In this case, the inevitable variation of activation above and below the mean level for an attractor state would provide the basis for assigning high and low confidence ratings. The same idea could apply to any of the representations that could serve as evidence for which word was presented when there is no clear percept of a particular word being presented. When the level of perceptual signal is low, the cleaned up percept based purely on noise (or a different stimulus from that perceived as being there) may be just as good a percept as that produced by signal and noise. Treisman and Schmidt (1982) and Treisman and Souther (1986) found that with rapid presentation of visual stimuli, subjects often reported seeing stimuli that were illusory conjunctions of the features presented. Such illusory conjunctions were reported with high confidence and (based on subjects' verbal reports) had the character of perceptual experiences. In such an experiment, confidence and accuracy could fail to correlate even when the subject sincerely reports consciously seeing.

The postulated top-down cleaning up of percepts is consistent with the fact that people easily confuse imagination and real stimuli of weak intensity (e.g. Perky, 1910, cited in Kelley & Rhodes, 2002; Wickless & Kirsch, 1989). Top down influences are likely to be particularly important the less time the subject has to appraise the stimulus. Lewicki & Czyzewska (2001) found that about 30% of undergraduates had split-second hallucinations "very often", e.g. they might have the

illusion of seeing an animal moving off the road, only to find out a moment later it was a piece of newspaper. A further 45% of students had this sort of experience "sometimes". In a subliminal perception experiment, thinking about or expecting a particular word, based on partial evidence provided by a different stimulus, might induce an active visual representation of that word just as good as that produced by the actual word itself. The activation of such errors could range to just as high a level as that of accurate representations of very impoverished stimuli. Mapping (2) could then be at chance, and the zero-correlation criterion would suggest perception had been subliminal when, in fact, whatever the subject saw (correctly or incorrectly) was seen quite consciously. The possible influence of expectation interfering with mapping (2) could be tested by relating individual differences in top down influences (e.g. as measured by Lewicki & Czyzewska, 2001) to the confidence-accuracy correlation.

This is an issue that deserves further investigation. Maybe when subjects give the wrong answer it was because the first order visual information systematically pointed in the wrong direction, but the system had no way of knowing this, no way of calibrating mapping (2). In this case, a zero-confidence accuracy correlation would not indicate the presence of unconscious mental states. Or maybe people were simply not aware of the perceptual evidence used in making the discriminations, i.e. perception was unconscious. This will be a matter for future research to determine. One approach is to derive predicted dissociations that more plausibly derive from the difference between conscious and unconscious mental states rather than between mental states that are systematically wrong rather than systematically right. In general, any measure of the conscious status of mental states shows its usefulness in participating in such theory driven research (Merikle, 1992). In fact, it is hard to see

how there could be any theory independent measure of the conscious or unconscious status of mental states.

Conclusion

Higher order theory, like Rosenthal's (1986) higher order thought theory, provides a tool by which we can see clearly the relevance of subjective rather than objective measures of being consciously aware of the world, and also analyse the appropriate use of various subjective measures such as the zero correlation criterion. Certain preconditions must be met before the zero correlation criterion can be applied in either subliminal perception or implicit learning paradigms as the confidence accuracy relationship depends not just on the strength of the mapping between the properties of first order mental states and the content of second order thoughts. Fortunately, one can often get a handle on the other relevant mappings, and hence plausibly use the zero-correlation criterion in implicit learning and perception research.

In this chapter we have considered the use of the zero correlation criterion where first order attitude is assessed by accuracy. Figure 1 indicates another possible measure of first order attitude, namely consistency. That is, another version of the zero correlation criterion – not considered in this chapter – is to measure the relationship between consistency and confidence ratings to determine if higher order thoughts (assessed by confidence ratings) are related to attitudes (assessed by consistency), and hence determine the conscious status of knowledge states. Lau (2002) showed that sometimes training on a dynamic control task led to dissociations

between confidence and consistency. He argued that this reflected the use of implicit knowledge. But ultimately the worth of the zero correlation criterion, in whatever guise it is used, is shown by participating in theory driven research: The criterion should separate knowledge types that have the different properties predicted by a psychological theory of the functioning of conscious and unconscious knowledge.

We have assumed in this chapter that the process of verbally reporting a second order thought is not problematic. In fact, as well as the three mappings considered in this chapter, there is a fourth mapping to consider, that is more or less fallible, and that is the mapping between the second order thought and verbal report. Dienes (in press) considers this mapping in detail. These four mappings indicate four ways in which verbal report can fail: It could fail because the content of the first order state is incorrect about the world (mapping 1, so a verbal report would fail to describe the world correctly); or because the first order attitude is inappropriate for the degree of correctness of the first order state (mapping 2, so reports of confidence would fail to relate to accuracy); or because a second order thought misrepresents the first order state (mapping 3, so verbal reports would fail to reflect actual first-order attitudes held); or because the verbal reports misrepresent the second order thought (so verbal reports would fail to reflect conscious experience).. This chapter, together with Dienes (in press), indicates how nonetheless verbal reports can be essential, if fallible, guides to determining whether a subject is consciously aware or merely aware of a stimulus or regularity.

References

Allwood, C. M., Granhag, P. A., Johansson, H. (2000). Realism in confidence judgements of performance based on implicit learning. European Journal of Cognitive Psychology, 12, 165-188.

Armstrong, D. (1980). The nature of mind and other essays. Cornell University Press.

Bjőrkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The under-confidence phenomenon. Perception & Psychophysics, 54, 75-81.

Block, N. (2001). Paradox and cross purposes in recent work on consciousness. Cognition, 79, 197-219

Carruthers, P. (1992). Consciousness and concepts. Proceedings of the Aristotelian Society, Supplementary Vol. LXVI, 42-59.

Carruthers, P. (2000). Phenomenal consciousness naturally. Cambridge: Cambridge University Press.

Chan, C. (1992). Implicit cognitive processes: theoretical issues and applications in computer systems design. Unpublished D.Phil thesis, University of Oxford.

Channon, S., Shanks, D., Johnstone, T., Vakili, K., Chin, J., & Sinclair, E. (2002). Is implicit learning spared in amnesia? Rule abstraction and item familiarity in artificial grammar learning. Neuropsychologia, 40, 2185-2197.

Chan, C. (1992). Implicit cognitive processes: theoretical issues and applications in computer systems design. Unpublished D.Phil thesis, University of Oxford.

Cleeremans, A. (1993). Mechanisms of Implicit Learning: Connectionist Models of Sequence Processing. Cambridge, MA: MIT Press

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. Trends in Cognitive Sciences, 2, 406-415.

Dienes, Z. (submitted). Assumptions of subjective measures of unconscious mental states: Higher order thoughts and bias.

Dienes, Z., & Altmann, G. (1997). Transfer of implicit knowledge across domains? How implicit and how abstract? In D. Berry (Ed.), How implicit is implicit learning? (pp 107-123). Oxford: Oxford University Press.

Dienes, Z., Altmann, G., Kwan, L, Goode, A. (1995) Unconscious knowledge of artificial grammars is applied strategically. Journal of Experimental Psychology: Learning, Memory, & Cognition, 21, 1322-1338.

Dienes, Z., & Berry, D. (1997). Implicit learning: below the subjective threshold. Psychonomic Bulletin and Review, 4, 3-23.

Dienes, Z., & Fahey, R. (1995). The role of specific instances in controlling a dynamic system. Journal of Experimental Psychology: Learning, Memory, & Cognition, 21, 848-862.

Dienes, Z., & Fahey, R. (1998) The role of implicit memory in controlling a dynamic system. Quarterly Journal of Experimental Psychology, 51A, 593-614.

Dienes, Z., Kurz, A., Bernhaupt, R., & Perner, J. (1997). Application of implicit knowledge: deterministic or probabilistic? Psychologica Belgica, 37, 89-112.

Dienes, Z., & Perner, J. (1999) A theory of implicit and explicit knowledge. Behavioural and Brain Sciences, 22, 735-755.

Dienes, Z., & Perner, J. (2001). When knowledge is unconscious because of conscious knowledge and vice versa. Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society, 1-4 August, Edinburgh, Scotland. Lawrence Erlbaum Associates: Mahwah, NJ (pp. 255-260).

Dienes, Z., & Perner, J. (2002a) A theory of the implicit nature of implicit learning. In French, R M & Cleeremans, A. (Eds), Implicit Learning and Consciousness: An Empirical, Philosophical, and Computational Consensus in the Making? Psychology Press.

Dienes, Z., & Perner, J. (2002b). The metacognitive implications of the implicit-explicit distinction. In Chambres, P., Izaute, M., & Marescaux, P.-J. (Eds),

Metacognition: Process, function, and use. Dordrecht, Netherlands: Kluwer Academic Publishers (pp 241-268).

Dienes, Z., & Perner, J. (1993). Unifying consciousness with explicit knowledge. In Cleeremans, A. (Ed.) The unity of consciousness: binding, integration, and dissociation. Oxford University Press.

Dulany, D.E., Carlson, R., & Dewey, G. (1984). A case of syntactical learning and judgement: How conscious and how abstract? Journal of Experimental Psychology: General, 113, 541-555.

Gigerenzer, G. (2000). Adaptive thinking: Rationality in the real world. Oxford: Oxford University Press.

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. Behavioural and Brain Sciences, 9, 1-66

Kelley, C. M., & Rhodes, M. G. (2002). Making sense and nonsense of experience: Attributions in memory and judgment. The Psychology of Learning and Motivation, 41, 293 - 320.

Kelley, S. W., Burton, A. M., Kato, T., & Akamatsu, S. (2001). Incidental learning of real world regularities in Britain and Japan. Psychological Science, 12, 86-89.

Kolb, F. C., & Braun, J. (1995). Blindsight in normal observers. Nature, 377, 336-338.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. Psychological Review, 100, 609-639.

Koriat, A. (1997). Monitoring one's knowledge during study: A cue-utilization approach to judgments of learning. Journal of Experimental Psychology: General, 126, 349-370.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and Accuracy of Near-Threshold Discrimination Responses. Consciousness and Cognition, 10,. 294-340.

Lau, K. K. (2002). Metacognitive measures of implicit learning in a dynamic control task. Unpublished D.Phil thesis, University of Sussex.

Lewicki, P., & Czyzewska, M. (2001). Styles of nonconscious intelligence. In Lewicki, P. & Cleeremans, A. (Ed.) Proceedings of the AISB'01 Symposium on Nonconscious Intelligence: From Natural to Artificial, 21st-24th March 2001, University of York. York, UK: University of York. (pp 43 –50.)

Merikle, P. M. (1992). Perception without awareness: Critical issues. American Psychologist, 47, 792-795.

Merikle, P. M., & Daneman, M. (1998). Psychological investigations of unconscious perception. Journal of Consciousness Studies, 5, 5-18.

Millikan, R. G. (1984). Language, thought, and other biological categories. Cambridge, MA: MIT Press.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing judgments. Psychological Bulletin, 95, 109-133.

Newell, B. R., & Bright, J. E. H. (2002). Well past midnight: Calling time on implicit invariant learning? European Journal of Cognitive psychology, 14, 185-205.

Peirce, C. S., & Jastrow, J. (1884). On small differences in sensation. Memoires of the National Academy of Sciences, 3, 73-83.

Reber, A.S. (1967). Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behaviour, 6, 855-863.

Reber, A.S. (1989). Implicit learning and tactic knowledge. Journal of Experimental Psychology: General, 118, 219-235.

Reber, A.S. (1993). Implicit learning and tacit knowledge: An essay on the cognitive unconscious. New York: Oxford University Press.

Reder, L.M. & Ritter, F. (1992) What determines initial feeling of knowing? Familiarity with question terms, not with the answer. Journal of Experimental Psychology:Learning, Memory, and Cognition18, 435-451.

Rosenthal, D.M. (1986). Two concepts of consciousness. Philosophical Studies, 49, 329-359.

Rosenthal, D.M. (2000).Consciousness, Content, and Metacognitive Judgments, Consciousness and Cognition, 9, 203-214.

Rosenthal, D. M (forthcoming).. "Consciousness and higher order thought", Encyclopedia of Cognitive Science. Macmillan.

Shanks, D. R. & St. John, M. F. (1994). Characteristics of dissociable human learning systems. Behavioural and Brain Sciences, 17, 367-448.

Siegal, S., & Castellan, N. J. (1988). Nonparametric statistics for the behavioural sciences, 2nd edition.McGraw Hill: London.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. Cognitive Psychology, 14, 107-141.

Treisman, A., & Souther, J. (1986). Illusory words: The roles of attention and of top-down constraints in conjoining letters to form words. Journal of Experimental Psychology: Human Perception and Performance, 12, 3-17.

Tunney, R. J., & Altmann, G. T. M. (2001). two modes of transfer in artificial grammar learning. Journal of Experimental Psychology: Learning, Memory, & Cognition, 27, 1322-1333.

Tunney, R.J. and Altmann, G.T.M. (1999) The transfer effect in artificial grammar learning: Re-appraising the evidence on the transfer of sequential dependencies. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25, 1322-1333.

Twyman, M. (2001). Metacognitive measures of implicit knowledge. Unpublished D.Phil thesis, University of Sussex.

Weiskrantz, L. (1988). Some contributions of neuropsychology of vision and memory to the problem of consciousness. In A. J. Marcel & E. Bisiach (Eds.), Consciousness in contemporary science (pp. 183-199). Oxford: Clarendon Press.

Weiskrantz, L. (1997). Consciousness lost and found. Oxford University Press.

Whittlesea, B. W. A., Brooks, L. R., & Westcott, C. (1994). After the learning is over: Factors controlling the selective application of general and particular knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 259-274.

Wickless, C., & Kirsch, I. (1989). The effect of verbal and experiential expectancy manipulations on hypnotic susceptibility. Journal of Personality and Social Psychology, 57, 762-768.

Appendix 1

Factors influencing mapping (2).

The aim of this appendix is to consider factors affecting the strength of mapping (2), the relation between accuracy and attitude strength. We will consider two measures of mapping (2), namely, Chan's difference score and the Goodman-Kruskal Gamma statistic. This is for purely illustrative reasons; the Chan difference score has often been used as a measure of the confidence-accuracy relationship in previous studies using the logic of the zero-correlation criterion, reviewed in the chapter; and Gamma, though rarely used in the implicit learning literature (for an exception, see Twyman, 2001), is the statistic of choice in the metacognition literature (Nelson, 1984).

As a measure of the attitude-accuracy relationship, the Chan difference score is the average attitude strength when a correct decision has been made minus the average attitude strength when an incorrect decision has been made. The higher score, the more positive the relation between attitude strength and accuracy. Gamma is a measure of association between two ordinally-scaled variables, each with two or more values (see Siegal & Castellan, 1988, pp 291-298).

Consider a set of strings, and towards each string the subject has a certain attitude that the string is grammatical or non-grammatical. We will assume the attitude is reflected perfectly in the consistency with which a subject gives a response. For example, if a subject responds "grammatical" to the string 80% of the time and "non-grammatical" 20% of the time, then the subject has an attitude strength of 0.8 that the string is grammatical. We can equivalently say the subject has an attitude

strength of 0.2 that the string is non-grammatical. We will first consider the Chan difference score, and then Gamma.

Let us say the accuracy over all the strings is 60%. Such an overall score could come about in various ways. For example, one way is by the subject having an attitude strength of 0.60 towards the correct response for each string individually. Thus, if there were 10 strings in total, we would expect there to be six correct decisions and four incorrect ones. For the correct decisions, the attitude strength would be 0.60 (the response is emitted with a probability of 0.60 for that item). For the incorrect decisions the attitude strength would be 0.40 (the response is emitted with a probability of 0.40 for that item).. So the Chan difference score would be 0.60 - 0.40 = 0.20. On this model, if the overall percent correct is labelled PC, then the probability of giving the right response for each item comes from a set with only one member, {PC}, e.g. {0.6}.

On another model, the subject gives each response with complete certainty, i.e. there is deterministic responding. Thus, the probability of giving the right response to each item comes from the set {0, 1}. For 10 items and an overall percent correct of 0.6, the subject would respond correctly to six items with an attitude strength of 1.0, and incorrectly to four items with an attitude strength of 1.0. The Chan difference score would be zero.

According to Reber (1989), subjects either guess the response to an item randomly, or they know the response to that item and respond perfectly consistently. Thus, on the Reber model, the probability of giving the right response to each item comes from the set {0.5, 1}. For example, for 10 items, if the subject guessed randomly for 8 of them, one would expect four to be correct (with an attitude strength of 0.5) and four incorrect (with an attitude strength of 0.5). By knowing the response

to the other two items (attitude strength of 1.0), there would be six items correct in total out of 10. The average attitude strength for the correct items would be (4*0.5 + 1+ 1)/6 = 0.67. The average attitude strength for the incorrect items is 0.5. The Chan difference score is 0.17.

Finally, one can consider models with the probability correct for each item drawn from some other set. We consider all the models described above, plus a model in which the probabilities are drawn from the set {0.1, 0.9}.

Figure A1 below shows the Chan difference score for mapping 2 for different probabilities correct. There are four curves, each created by assuming the probability, p, of a correct response to any item could only be selected from a given specified set. For example, on one extreme, the probability of correct response to each item was just the overall percent correct over all items. This is the top line. Here is no variance in p's between items for a given percent correct. At the other extreme, the probability of correct response to each item was either 1 or 0, there is maximum variance between the p's across items. This is the bottom thick line.

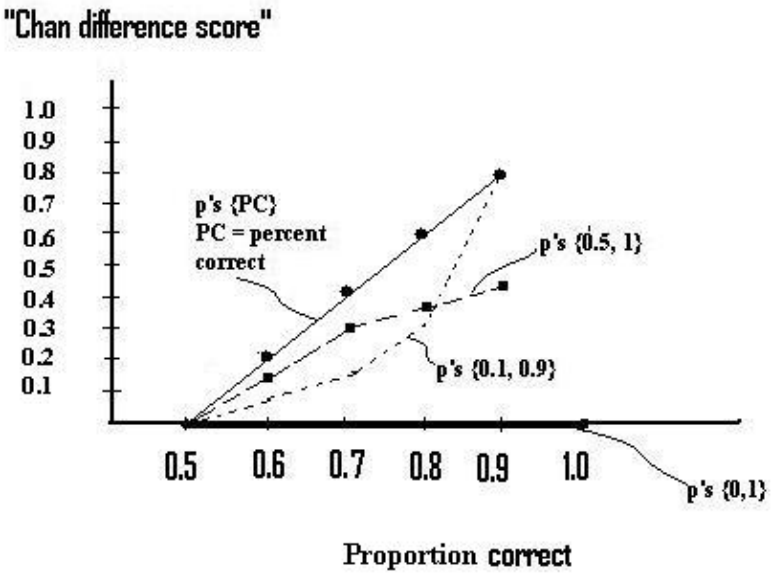


Figure A1
The Chan difference score measure of mapping 2 for different distributions of attitudes over items.

Note that the Chan difference score is not defined for an overall proportion correct of 1.0 because there are no incorrect answers. For a proportion correct of 0.9, the {0.1, 0.9} model becomes a degenerate case, because all strings must have a proportion of being correct of 0.9. Thus, the {0.9, 0.1} model is the same as the {PC} model in this case.

For the bottom curve, the Chan difference score is always zero. Mapping 2 is zero and so the assumptions of the zero-correlation criterion are not satisfied. (This is the case mentioned in the text, page *15*, where EC=0.) In no case is the Chan difference score negative. Mapping (2) is always positive or else zero. Is there any indicator of the strength of the mapping (2)? Consider a test where subjects are tested on each item twice. Figure A2 below shows how the proportion of times an item would be classified twice in error (EE) for each model and percentage correct.

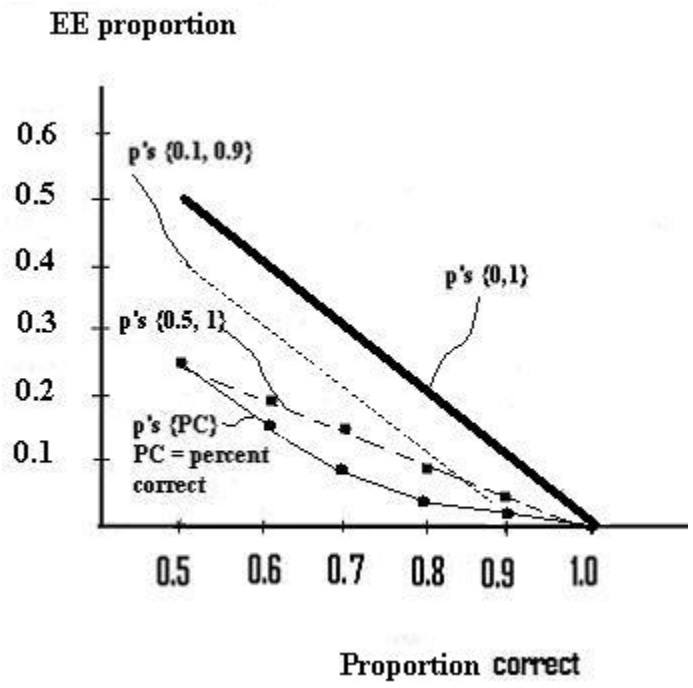


Figure A2
A measure of proportion of items classified in error twice in a row (EE) for different distributions of attitudes over items.

For a given percent correct, the ranking of the Chan difference score, from highest to lowest, is always the same as the ranking of EE scores from lowest to highest. In other words, for a given percent correct, the lower the EE proportion, the better is mapping (2), as assessed by the Chan difference score. This is useful in comparing two conditions with the same percent correct; if one has a higher EE than the other, it is reasonable to conclude mapping (2) is worse. This fact is used in the text to draw conclusions about the relative amount of unconscious knowledge in two conditions, as explained in the text (p *19*).

Now we will consider another measure of mapping 2, Gamma. All the models we are considering here generate just two attitudes (e.g. the {0.5, 1.0} model involves just attitude strengths of either 0.5 or 1.0), so the relevant data for calculating Gamma can be represented by a 2 X 2 table, where a, b, c, and d are counts:

	attitude1	attitude 2
correct	a	b
incorrect	c	d

The formula for Gamma reduces to (ad-bc)/(ad+bc) for the 2 X 2 case. Gamma varies between +1 and -1, where 0 indicates no association.

Nelson (1984) argued that Gamma had various properties desirable in a measure of metacognition; for example, if a subject has perfect metaknowledge this can be reflected in a Gamma of 1, regardless of the overall level of performance (whereas the Chan difference score depends on this level); it does not assume an interval scale for either variable (in contrast, the Chan difference score assumes an interval scale); and it is not sensitive to ties (in the general n X n case, in contrast to other nonparametric measures of associations, like Kendall's Tau).

Figure A3 below shows the Gamma for each model considered above. Note that for the {0,1} model, Gamma is undefined or zero, as there is only one attitude. The other models are displayed on the figure. For the {PC} and {1, 0.5} models, Gamma is uniformly one (except for a percent correct of 0.5 and 1, where there is only one attitude). This is unlike the Chan difference score, which increased with percent correct for these models. For the {0.1, 0.9} model, Gamma does increase with percent correct, like the Chan difference score. Dienes, Kurz, Bernhaupt, and Perner (1997) argued that the pattern of subjects' responding over repeated testing of the same items was most consistent with a model in which each subject responded with a range of probabilities correct for different items (and which are different for different subjects), not just {1, 0.5, PC}; that is, one might expect in general the underlying Gamma measure of attitude-accuracy association to increase with percent correct in the artificial grammar learning paradigm.

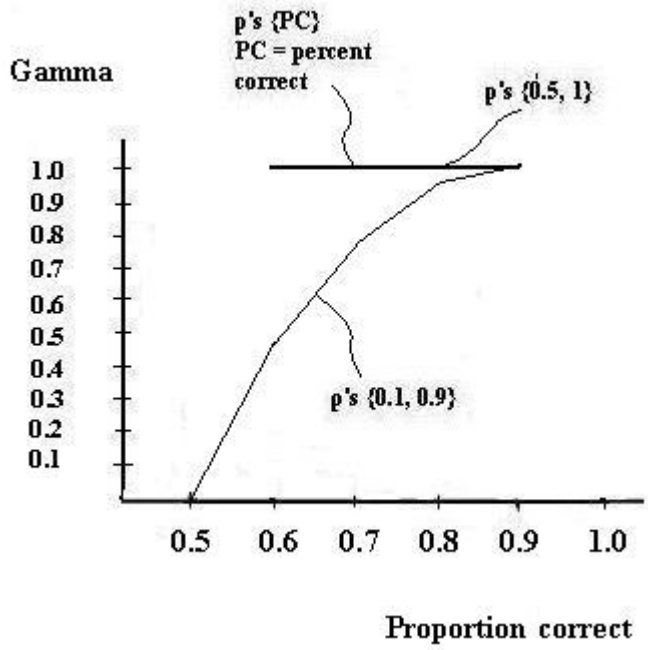


Figure A3

The Gamma measure of mapping 2 for different distributions of attitudes over items

Where the Gamma measure of mapping (2) does vary for the same percent correct between different models, the model with the higher EE has the lower measured strength of mapping (2) (as was the case for the Chan difference score).

In summary, we have considered two measures of the strength of mapping (2), the Chan difference score and Gamma. Where the subject's first-order attitude is not certainty for every item, and overall classification performance is between 0.5 and 1.0, both produce positive measures of mapping (2) for the models considered. For both measures, EE can be used as an index of the strength of mapping (2) for the same overall classification performances. Gamma has some nicer properties than the Chan difference score, but the Chan difference score still produces interpretable results (i.e. for the same percent correct, the measured strength of mapping (2) varies inversely with EE, a property used in the text, p *19*).

If mapping (3) was perfect, the Chan difference score and Gamma calculated on confidence ratings, rather than first-order attitudes, would vary according to overall percent correct in the way that mapping (2) expresses itself in these measures, as explored in this appendix.

Finally, we note that subjects' apparent use of probability matching in expressing first-order attitudes is essentially irrational. If I believe a string is grammatical with 80% certainty, rationally I should say "grammatical" on every occasion. But subjects do not; sometimes they say "non-grammatical", an answer about which they are fairly certain is wrong ("fairly certain" in the sense of a first-order attitude). Presumably, on those trials it does not seem to the subject that they believe they are wrong, otherwise they would not give that answer; that is, presumably their HOTs do not match their attitudes. If subjects were perfectly aware

of their attitudes, either they would sometimes give a response about which they have a confidence of less than 50%, or more likely, they would not probability match at all. If subjects have partial awareness of their attitudes, their HOTs may always lie in the range 50-100 even while the subject probability matches, but nonetheless, stronger attitudes may be associated with higher confidence ratings on average. If subjects have no awareness of their attitudes (the knowledge is unconscious), confidence ratings will bear no relation to attitudes.