Available online at www.sciencedirect.com
ScienceDirect
Consciousness and Cognition 15 (2006) 673–692
Consciousness and Cognition
www.elsevier.com/locate/concog

How something can be said about telling more than we can know: On choice blindness and introspection

Petter Johansson, Lars Hall *, Sverker Sikström, Betty Tärning, Andreas Lind

Lund University Cognitive Science (LUCS), Lund University, Kungshuset, Lundagård, 222 22 Lund, Sweden

Received 12 June 2006
Available online 17 October 2006

Abstract

The legacy of Nisbett and Wilson's classic article, *Telling More Than We Can Know*: *Verbal Reports on Mental Processes* (1977), is mixed. It is perhaps the most cited article in the recent history of consciousness studies, yet no empirical research program currently exists that continues the work presented in the article. To remedy this, we have introduced an experimental paradigm we call choice blindness [Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–119.]. In the choice blindness paradigm participants fail to notice mismatches between their intended choice and the outcome they are presented with, while nevertheless offering introspectively derived reasons for why they chose the way they did. In this article, we use word-frequency and latent semantic analysis (LSA) to investigate a corpus of introspective reports collected within the choice blindness paradigm. We contrast the introspective reasons given in non-manipulated vs. manipulated trials, but find very few differences between these two groups of reports.

Keywords: Introspection; Verbal report; Confabulation; Choice blindness; Change blindness; Word-frequency analysis; Latent Semantic Analysis

1. Introduction

Nearly, thirty years have passed since the publication of Nisbett and Wilson's seminal article *Telling More Than We Can Know: Verbal Reports on Mental Processes* (1977). Arguably, this article is one of the most widely spread and cited works on the nature of introspection ever to be published. As of May 2006, according to the ISI Web of Science Index, Nisbett and Wilson (1977) have been cited an astonishing 2633 times.[1]

* Corresponding author. Fax: +46 46 222 4424.
E-mail address: Lars.Hall@lucs.lu.se (L. Hall).

[1] To put these numbers in perspective it is more than five time as many citations as that gathered by Thomas Nagel's classic essay "*What is it like to be a bat?*" (1974), nearly ten times as many as that given to any of Benjamin Libet's famous articles on the subjective timing of conscious will, and more than twice as many as the combined cites given to *all* the articles that have appeared in the *Journal of Consciousness Studies* and in *Consciousness and Cognition* during the last ten years.

1053-8100/$ - see front matter
doi:10.1016/j.concog.2006.09.004

No doubt there are many reasons for these extraordinary citation numbers. The comprehensive and accessible review of N&W has long held an attraction for applied researchers dealing with different forms of verbal report. These citations come from the most diverse fields of research: nursing studies, human–computer interface design, demography, psychotherapy, sports psychology, etc.[2] More specifically, N&W has become part of the ''checks and balances'' of survey and consumer research, as a basic item that must be considered, like experimental demand effects, or the possibility of sampling error (Schwarz & Oyserman, 2001).

Yet, despite this, no systematic empirical research program exists that carry on the pioneering work of N&W. It is a piece everybody seems to return to, but hardly anybody tries to improve upon. Buried in the mass of citations one can find a group of articles from the eighties that strove to advance the methodology of N&W (see, e.g., Guerin & Innes, 1981; Morris, 1981; Quattrone, 1985; Sabini & Silver, 1981; Sprangers, Vandenbrink, Vanheerden, & Hoogstraten, 1987), but the output from this initiative is all but invisible in the current debate. Despite the prolific work of Wilson himself, who has taken the general idea of lack of introspective access in several new directions (e.g., Wilson, 2002; Wilson & Kraft, 1993; Wilson, Laser, & Stone, 1982; Wilson, Lindsey, & Schooler, 2000), the empirical debate about N&W soon came to a standstill, with multiple layers of inconclusiveness confusing just about everyone involved (as meticulously summarized by White (1988) in his tenth anniversary review of N&W).

Consequently, then, when a scholarly reviewer like Goldman (2004) discusses the epistemic status of introspective reports, he feels the need to address (and refute) the 27-year-old ''*challenge from Nisbett and Wilson*,'' rather than some red-hot contemporary alternative.

It is ironic that the exemplary structure of the original article might be partly to blame for this lack of development. N&W not only tried to show experimentally that ''there may be little or no direct access to higher order cognitive processes'' (1977, p. 231), but they also tried to present an explicit framework for future studies, and a fully fledged alternative theory about the origins of introspective reports (thereby taking upon themselves a burden of explanation that most researchers would shun like the plague).[3] Their basic idea was that the accuracy of introspective reports could be determined by comparing the reports of participants in the experiments to those of a control group who were given a general description of the situation and asked to predict how the participants would react—the so-called *actor–observer* paradigm (Nisbett & Bellows, 1977). If actors consistently gave more accurate reports about the reasons for their behavior than observers did, then this would indicate privileged sources of information underlying these reports. If not, then the position of N&W would be further supported.

Unfortunately, as is shown by the contributions of White (1988) and others (e.g., Gavanski & Hoffman, 1986; Kraut & Lewis, 1982; Wilson & Stone, 1985; Wright & Rip, 1981), it is an exceedingly complex task to unravel all the possible influences on report in an actor–observer paradigm (and this was *before* the whole simulation vs. theory–theory debate got started, which complicates things even further, see Rakover (1983) for an early hint of this debate to come). White (1987) writes:

> In [its] original form the proposal [of N&W] foundered, largely because it is at present untestable. It is difficult if not impossible to ascertain the nature and extent of involvement of ''introspective access,'' whatever that is, in the generation of causal reports, and one cannot assume a straightforward relationship between ''introspective access'' and report accuracy. In addition, a valid distinction between ''process'' and ''content'' or ''product'' has yet to be pinned down, despite some attempts to do so. Given these problems, the proposal effectively degenerated into a simpler hypothesis that causal report accuracy cannot be significantly enhanced by information about relevant mental activity between stimulus and response. As we have seen, tests of this hypothesis have so far proved inconclusive. But to continue refining such tests with the aspiration of good internal validity is likely to prove an empty methodological exercise (p. 313).

[2] See for example Brewer, Linder, Vanraalte, and Vanraalte, 1991; Higuchi and Donald, 2002; Jopling, 2001; Jorgensen, 1990; Sandberg, 2005.

[3] It would seem incumbent on one who takes a position that denies the possibility of introspective access to higher order processes to account for these reports by specifying their source. If it is not direct introspective access to a memory of the processes involved, what is the source of such verbal reports? (Nisbett & Wilson, 1977, p. 232).

Thus, with an initially promising but ultimately too narrow conception of how to refine the N&W approach, this line of empirical investigation of introspection ground to a halt. While the disillusioned quote from White might suggest a more general point, that empirical studies of introspection will always be subjected to wildly differing conceptual analyses (of "content", "access", "process", etc.), and that no amount of empirical tinkering is likely to satisfy the proponents of the different consciousness camps (Rorty, 1993), we do not share this gloomy outlook. In our view, the lacuna left in the literature after the collapse of the actor–observer paradigm ought to be seen as a challenge and an invitation. After almost thirty years of intensive research on human cognition, it really *ought* to be possible to improve upon the experimental design of Nisbett and Wilson (1977).

2. Choice blindness and introspective report

In Johansson, Hall, Sikström, and Olsson (2005), we showed that participants may fail to notice mismatches between intention and outcome when deciding which face they prefer the most. In this study participants were shown pairs of pictures of female faces, and were given the task of choosing which face in each pair they found most attractive. In addition, on some trials, immediately after the choice, they were asked to verbally describe the reasons for choosing the way they did (the participants had been informed in advance that we would solicit verbal reports about their intentions during the experiment, but not the specific trials for which this was the case). Unknown to the participants, on certain trials, a double-card ploy was used to covertly exchange one face for the other. Thus, on these trials, the outcome of the choice became the opposite of what they intended.

We registered both concurrently and in post-test interviews whether the participants noticed that anything went wrong with their choice. Tallying across all the different conditions of the experiment, no more than 26% of all manipulation trials (M-trials) were exposed. We call this effect *choice blindness* (for details, see Johansson et al., 2005).

To solicit the verbal reports we simply asked the participants to state *why* they chose they way they did. As Nisbett and Wilson (1977) remarked in the opening lines of their article: "In our daily life we answer many such questions about the cognitive processes underlying our choices, evaluations, judgments and behavior" (p 231). Thus, for the non-manipulated trials (NM-trials) we expected straightforward answers in reply. For the M-trials, on the other hand, the situation was very different. Here, we asked the participants to describe the reasons behind a choice they did not in fact make. Intuitively, it is difficult to envisage how one would respond to such an anomaly (i.e., we simply do not know what it is like to say why we prefer a particular picture, when we in fact we chose the opposite one). But based on common sense alone, one would suspect that the reports given for NM- and M-trials would differ in many ways.

To explore this contrast, we identified three main psychological dimensions that we believed could be used to differentiate between the reports given in response to NM- and M-trials. These dimensions concerned the *emotionality*, *specificity*, and the *certainty* of the reports. Our reasoning was that participants responding to a manipulated face ought to show less emotional engagement, as this was actually the alternative they did not prefer (*emotionality*); they also ought to make less specific and detailed reports, as no prior reasons have been formulated for this alternative (*specificity*); and they ought to express less certainty about their choice (*certainty*). As detailed in Johansson et al. (2005), we found no differences between the NM- and M-reports on these three dimensions.

In our view, these unexpected commonalities between NM- and M-reports raise many interesting questions about the nature of introspection. However, before any attempts to relate this result to current theories of consciousness are made, we believe the contrastive methodology as such needs to be further discussed and refined.

Debates about the validity and reliability of introspective report often involve lots of back and forth on clinical syndromes where confabulation is likely to be found (such as split-brain, hemineglect, hysterical blindness, or Korsakoff's syndrome, e.g., see Hirstein, 2005). What is striking about these cases is that the patients say things that are severely disconnected from everyday reality. The reports may not always be fantastic or incoherent, but we can easily check the state of the world and conclude that they are implausible as candidate explanations of their behavior. However, as confabulation is defined in contrast to normality, we run into problems when trying to investigate the mechanisms behind the phenomenon. As the confusion and stalemate on Nisbett and Wilson's actor–observer paradigm demonstrates, without the benefit of good contrast cases to work from, discussions of the possibility of confabulatory reporting in normal human populations tend to take on a distressingly nebulous form. The position of N&W was essentially that there are elements of confabulation in all introspective reports, but that these confabulations nevertheless are plausible and reasoned (based on either shared cultural beliefs or idiosyncratic theorizing). But how do we go about testing this interesting proposition, if we cannot even determine what a "genuine" introspective report should look like?

It is our hope that the analysis of introspective reports in our choice-blindness paradigm can contribute toward the goal of establishing a better grip on what constitutes truthful and confabulatory report, and to discern interesting patterns of responding along this dimension with respect to both individual variation and the context of choice.

In Johansson et al. (2005), to compare and contrast the NM- and M-conditions we used blind independent raters to evaluate each of the reports (thus following the natural instinct of experimental psychologists to ground any exploratory measurements by the concept of interrater agreement). But this is not the only way to conduct such an investigation. An obvious weakness of relying on naïve raters to refine the categories used is that they might fail to discern possible differences in the material that could have been revealed by expert analysis. In addition, on the flip side, there is a problem of potential bias in our original choice of categories. Who are *we* to decide what constraints that can be made on the potential contrasts between the NM- and the M-reports?

Thus, in this article, using a new corpus of introspective reports, we present two additional approaches to the same task. First, we carry out an expert-driven linguistic analysis based on word-frequency counts. This analysis covers a great range of linguistic markers known to be important for contrasting different text corpuses, and functions as a complementary top-down way of capturing and recreating the psychological dimensions used in Johansson et al. (2005) (see description above). But while these dimensions are bound to be a reflection of the folk-psychological invariance of everyday life (i.e., everybody has experienced differing degrees of uncertainty and emotionality, etc.), we should be open to the possibility that a computational cognitive perspective might settle on far less intuitive contrasts as being the most productive for analyzing this type of material. To this end, as a more exploratory and data-driven approach, we introduce a novel implementation of Latent Semantic Analysis (LSA). As LSA creates a multi-dimensional semantic space using very few theoretical assumptions, it is perfectly suited to investigate possible similarities and differences between the NM- and M-reports that cannot easily be captured with the standard toolkit of linguistic and psychological analysis.

3. The corpus of reports

The corpus of introspective reports used for our analysis was collected in a recent study extending our previous choice blindness results (Hall, Johansson, Tärning, & Sikström, in prepin press). As in Johansson et al. (2005), participants in this study were shown pairs of pictures of female faces, and were asked to choose which face in each pair they found most attractive. We constructed the face pairs in order to vary the discrepancy of attractiveness within each pair, while an attempt was made to keep similarity constant at an intermediate level (i.e., clearly different, but not drastically so, see Hall et al., in prep).

Each participant completed a series of 15 face-pairs, with four seconds of deliberation time given for each choice. As in the previous study, six of the pairs were designated as verbal report-pairs, and any three of these six were in turn manipulated for each participant. Eighty participants (49 female) took part in the study (mean age 24.1, *SD* 4.1), which gives a total of 480 reports collected.

The collection of introspective reports is rich and varied. For the reader to be able to get a descriptive feel for the contents of the reports, Table 1 shows an illustrative selection of statements from both the NM- and M-trials.

To find out the opinion of the participants about the study, we conducted a semi-structured post-test interview. The interview sessions revealed that a great majority of participants felt that the given task was interesting, and that four seconds was enough time to make a meaningful choice (however, there was also a great range and natural variability within the reports, with both self-assured enthusiasm, and concerned caution at times).

Table 1
Extracts from the NM- and M-reports

Non-manipulated	Manipulated
It was her eyes that struck me right away, they are so incredibly, ehh... awake, you might say... it looks as if they want to explore everything	She looked more pleasant, looks very kind, ehh [pause] reminds me of a friend that... a good friend of mine
Nice eyes [pause] neat haircut, neat hair... ehm [pause] well... she had a nice nose too...	hmm [pause] well the eyes were very big and beautiful, and it is often the eyes people look at, or at least, that's what I do
Evenly sized irises, an even sized radius for the irises and the pupils	There's a lot of cheeks there, and it looks soft and receptive and it's a generous nose too
The eyes are radiating there, and the mouth too, it has that little... about to smile thing going on	Well it is the eyes, I like big eyes... hmm... and then she's got a nice mouth, very shapely I think
I'm thinking that she is, that is, keen on the arts or something, that is, that is, an aesthetic... feeling	That was easier she looks much more alive, ehh... there's there's much more spark in her eyes
And this is a much more receptive face	No, I do not know, she, the other one had a more pointy chin, and so
Again, she was just more beautiful than she [pause] than the other one	ehh... I believe I think she had more atmosphere to her look, or whatever one might call it... ehm
The other one looked a bit crazy, I guess this one had a better nose	ehh, because [pause] she's more well kept may be
She looks a bit pale and frightened... looks like she is in a need of a vacation at the beach	A bit like this, nice you know, a bit wimpy [laughter]
Well, maybe the impression and not so much the details you know, and the way she looks	I believe it is because she looks a bit more, a bit special, I do not know if it is the hair or the shape of her face, I think, and so

Note. Extracts from the NM- and M-reports. The statements were chosen to display the range of responses present in the corpus, with examples taken from reports both high and low on one or more of the dimensions *specificity*, *emotionality*, and *complexity*. The extracts are taken from both the short and the long reports, with a rough matching on the three previously mentioned dimensions being made across the NM- and M-columns.

The overall detection rate for the manipulated trials was roughly equivalent to our prior results, with 27.5% of the trials detected (for details, see Hall et al., in prep). Adjusting for detections left 414 reports, and for technical reasons (mishap with the recorder, indecipherable talk, etc.) another 23 were omitted, which leaves 228 NM- and 163 M-reports for the final analysis.

In addition, the study was divided into two different conditions for the introspective reports. The first condition mirrored our previous setup, where we simply asked the participants to state the reasons for choosing the way they did. Here, interaction with the experimenter was kept at an absolute minimum, and no attempts were made to further prompt the participants once they spontaneously seceded in their talk. In the second condition, the same question was posed, but the experimenter encouraged the participants to elaborate their answers up to one full minute of talking time. This was done both by the use of positive non-verbal signals, such as nodding and smiling, and by their linguistic equivalents (such as saying "yes, yes"), and by interjecting simple follow-up questions (such as "what's more ?", or "what else did you think of ?"). The reason we included the second condition was to see whether longer reports would produce a clearer differentiation between NM- and M-trials.[4] The reports elicited in the first condition are referred to as short reports and reports from the second condition are referred to as long reports. The average length of the reports was 20 words for the short ones and 97 words for the long ones. All reports were recorded digitally, and later transcribed. The utterances of the experimenter were transcribed, but removed from the corpus before analysis. Pauses, filled hesitations, laughter, and interjections are included in the corpus, but were not counted as words when establishing relative word frequencies between the reports. The final number of reports included in the analysis calculated by condition was 111 (NM-short), 117 (NM-long), 81 (M-short), and 82 (M-long).

[4] This can be read both in the sense that the inclusion of more words in the study would increase the statistical power of the analysis, and that potentially confabulatory elements would be more prominent, making a possible contrast between the two types of report more vivid. It should be noted that this condition also served a role in the second focus of the study, which was to investigate whether choice might influence preference change (see Hall et al., in prep).

4. Comparative linguistic analysis

In linguistics, research is often concerned with examining structural differences between different corpora of spoken or written text. Typical examples include comparing different stages in the language development of children (Durán, Malvern, Richards, & Chipere, 2004), contrasting spoken and written text (Biber, 1988), or attempting to authenticate all the works named as Shakespeare's (Elliot & Valenza, to appear).

The methods used to establish such contrasts are diverse, but they all strive to find distinctive markers, a linguistic "fingerprint" that says something interesting about the text under study (Biber, 1988; Labov, 1972). When investigating psychological aspects of language use, emphasis is normally placed on contextual factors influencing the situation, such as the relative status between the speakers, the conversational demands inherent in the situation, and obviously the history and personality of the speakers involved (Brown & Yule, 1983; Norrby, 2004). But the pitfalls of this type of qualitative content analysis are well known (Krippendorff, 1980), and any form of interpretative approach becomes increasingly laborious and ungainly as the amount of text increases.

However, an accumulating body of evidence suggests that a great number of factors can be discerned by analyzing the overall frequency of words used in a text, even if it means ignoring the actual content of the sentences produced. Pennebaker and co-workers have developed a method to differentiate between two (or more) corpora by systematically counting the words used (Pennebaker, Mehl, & Niederhoffer, 2003). They have built a large-scale database consisting of weighted and validated categories, such as words related to cognition ("cause," "know"), emotion ("happy," "bitter"), space ("around," "above"), as well as standard linguistic types (articles, prepositions, pronouns). This database has then been implemented in a specialized program called Linguistic Inquiry and Word Counting (LIWC), which is capable of sifting and sorting all the words from a particular text into the above-mentioned categories, thereby creating a linguistic profile of the text under study (Pennebaker, Francis, & Booth, 2001). Using LIWC, they have managed to establish telling differences between texts for such diverse areas as suicidal and non-suicidal poets (Stirman & Pennebaker, 2001), Internet chat rooms the weeks before and after the death of Lady Diana (Stone & Pennebaker, 2002), and language change over the life span (Pennebaker & Stone, 2003).

While issues of translation from Swedish to English barred us from using the LIWC program on our corpus of reports, we were able to implement our own version of the same methodology using a combination of commercial programs (CLAN), and homemade scripts written to solve specific problems during the analysis. The basic procedure then, for most of our measures, was that we identified different types of words and categories of interest, and then established their relative frequency in the material. These relative frequencies (the occurrence of the target category divided by the total number of words for each report) are the main unit used when comparing NM- and M-reports. Unless otherwise stated, the statistic used is Mann–Whitney *U*-test. A non-paired non-parametric test is used as there is an unequal amount of NM and M trials (due to the removal of detected M-trials), and because most of the variables did not follow a normal distribution curve.

As we stressed in the introduction, the analysis performed in this article is largely exploratory. Choice blindness is a new experimental paradigm, and the best we have been able to get from the research literature is guiding hunches and intriguing leads about what factors should go into the analysis. Thus, the categorization of the results below should not be read as carving deep metaphysical divisions, but rather as an attempt at pedagogical clustering to highlight interesting patterns for the reader.

In the presentation the English translations always appear in italics, and the original Swedish sentences or words appear in the following parentheses. Unless specifically mentioned, all presented comparisons between the NM- and M-reports below include both the short and the long condition. For ease of reference we have included a summarizing table at the end of the section, with detailed numbers for all the measures used (see Table 2).

4.1. Uncertainty

The most obvious contrast to make between the NM- and M-reports concerns the degree of certainty expressed by the participants in their reports. In (Johansson et al., 2005), our blind raters felt that this was the easiest dimension to discern, and the one most firmly represented in the material. But this is not something

Table 2
Summary of the results from the contrastive linguistic analysis

	Short NM	Short M	*p*	Long NM	Long M	*p*
Six words marking uncertainty	0.060 (0.007)	0.065 (0.010)	0.999	0.036 (0.002)	0.039 (0.002)	0.438
Extended measure of uncertainty	0.096 (0.009)	0.101 (0.011)	0.728	0.071 (0.007)	0.077 (0.008)	0.105
Filled pauses	0.047 (0.006)	0.047 (0.006)	0.452	0.048 (0.003)	0.054 (0.004)	0.228
Unfilled pauses	0.018 (0.005)	0.036 (0.015)	0.135	0.032 (0.003)	0.041 (0.005)	0.262
Laughter	0.010 (0.003)	0.019 (0.005)	0.343	0.008 (0.001)	0.010 (0.002)	0.590
Metalingual comments	0.493 (0.032)	0.544 (0.035)	0.296	0.543 (0.017)	0.544 (0.019)	0.745
Nouns	0.091 (0.009)	0.078 (0.009)	0.348	*0.089 (0.003)*	*0.078 (0.004)*	*0.019*
Specific nouns	0.055 (0.008)	0.043 (0.008)	0.320	0.052 (0.003)	0.046 (0.003)	0.178
Non-specific nouns	0.029 (0.005)	0.022 (0.004)	0.604	0.025 (0.002)	0.020 (0.002)	0.103
Nouns (Johansson et al., 2005)	0.105 (0.009)	0.113 (0.011)	0.543	*	*	*
Specific nouns (Johansson et al., 2005)	0.056 (0.007)	0.069 (0.011)	0.310	*	*	*
Non-specific nouns (Johansson et al., 2005)	0.049 (0.007)	0.044 (0.006)	0.543	*	*	*
Adjectives	0.121 (0.009)	0.121 (0.009)	0.155	0.115 (0.004)	0.115 (0.004)	0.284
Adjectives (positive)	0.054 (0.008)	0.047 (0.007)	0.853	*0.047 (0.003)*	*0.037 (0.003)*	*0.016*
Adjectives (negative)	0.004 (0.002)	0.009 (0.003)	0.472	0.012 (0.001)	0.013 (0.002)	0.729
Adjectives(Johansson et al., 2005)	0.116 (0.008)	0.108 (0.008)	0.511	*	*	*
Adjectives (positive) (Johansson et al., 2005)	0.094 (0.008)	0.087 (0.008)	0.557	*	*	*
Adjectives (negative) (Johansson et al., 2005)	0.022 (0.003)	0.021 (0.003)	0.849	*	*	*
Word length	4.288 (0.745)	4.403 (0.916)	0.339	5.215 (0.614)	5.265 (0.579)	0.557
Lexical density	0.331 (0.014)	0.317 (0.013)	0.453	0.303 (0.005)	0.290 (0.006)	0.130
Lexical diversity	*	*	*	$D = 53.015$ (2.308)	$D = 49.528$ (2.089)	0.369
Priming. new nouns	1.144 (0.111)	1.086 (0.140)	0.483	3.701 (0.211)	3.744 (0.322)	0.424
WHY present	0.225 (0.040)	0.173 (0.042)	0.376	0.838 (0.034)	0.927 (0.029)	0.062
WHY past	0.162 (0.035)	0.086 (0.031)	0.125	0.393 (0.045)	0.317 (0.052)	0.274
COMP present	0.108 (0.030)	0.037 (0.021)	0.071	0.137 (0.032)	0.085 (0.031)	0.267
COMP past	0.315 (0.044)	0.407 (0.055)	0.190	0.453 (0.046)	0.585 (0.055)	0.066
First-person pronouns	0.071 (0.007)	0.081 (0.009)	0.676	0.047 (0.003)	0.053 (0.004)	0.191
Third-person pronouns	0.123 (0.006)	0.116 (0.009)	0.800	0.108 (0.003)	0.112 (0.004)	0.646
Tense. verbforms present	0.107 (0.008)	0.115 (0.013)	0.599	0.104 (0.004)	0.111 (0.004)	0.281
Tense. verbforms past	0.080 (0.008)	0.077 (0.009)	0.746	0.053 (0.003)	0.051 (0.004)	0.612

Note. The number in parentheses is the standard deviation of the mean. The italic sections represent the significant differences found between the NM- and M-reports. An asterisk denotes that the measurement was not applicable for this cell.

peculiar to our particular corpus. The study of certainty has a long history in contrastive linguistics. It has, for example, been argued that female language often contains more words expressing uncertainty, and that it often is more imprecise and non-committal (Lakoff, 1975). The argument is centered on distinctive markers of uncertainty, such as *sort of, I think*, and *you know*, a class of expressions and words called *hedges* (Holmes, 1995, 1997). Similarly, differences in expressed certainty have been found between different social classes, academic disciplines (Vartatala, 2001), and even within the same research fields when different languages are used (Vold, 2006). An issue closely related to hedging is *epistemic modality*, which concerns how we express our level of commitment to the propositions we produce. What is examined here is not just uncertainty but the full spectrum of security in a statement—from *I know it's true* to *I guess it's true* (Frawley, 1992).

However, when looking for markers of uncertainty, it is important to note that there are several different aspects of uncertainty at play in our material. First, the participants might be unsure about the decision, indicating that they do not know *why* they chose one face over the other. Second, they might be hesitant about the act of speaking itself, simply not knowing what to say next. Third, the participant might feel uncomfortable and cautious about the situation as such, sensing that something is wrong, but just not knowing what it is. Following the literature, we created several different measures to try to capture a very broad sense of uncertainty.

For the epistemic aspect of uncertainty, we set up a list of words and phrases with an established function as hedges: *perhaps* (kanske), *you know* (ju), *I suppose* (väl), *probably* (nog), *do not know* (vet inte), *I think* (tror jag). These particular hedges were chosen because they were highly frequent in our corpus, thus making them good candidates for being able to differentiate between the NM- and M-reports. For the calculations we used a composite measure based on the relative frequency of the class of hedges compared to all words for each report. This was done both as a group and for each individual word or phrase. However, we found no statistical differences between the NM- and the M-reports for epistemic uncertainty, neither for the short nor for the long condition.[5]

As a measure of hesitance, we used both filled and unfilled pauses in the speech. An unfilled pause was defined as a silence within sentences lasting for more than 0.5 s. The filled pauses consisted of vocalizations filling the gaps between words, as well as words without content or function in the linguistic context (e.g. um, er, *na* (nä), *yeah* (jo)). As such, pauses have been hypothesized to be an instrument for the speaker to manage his or her own cognitive and communicative processes—i.e., to buy time while planning what to say next (Allwood, 1998). Given the intuitive assumption about the choice blindness situation that the entirety of the verbal explanation is constructed on the spot, an analysis of pauses seemed to us to be a very promising measure to use. But as was the case for the epistemic markers, we found no significant differences between NM- and M-reports for the amount of pauses used. As an independent category of filler activity, we also calculated the amount of laughter present in the NM- and M-reports (the hypothesis being that laughter can function as a signal of nervousness, distress, or surprise, see Glenn, 2003), but again, we found no significant differences with respect to laughter between the NM- and M-reports.

In summary, using several different linguistic measures, we found no evidence of differences in expressed uncertainty between the NM- and M-reports.

4.2. Specificity

The crux of the dilemma in the choice blindness paradigm is what sources the participants draw upon, or what mechanisms they use, when delivering their introspective reports in the NM- and the M-trials. Again, the common-sense assumption would be that the NM-reports reflect the actual intention that resulted from the deliberation phase (this being a natural source of information when stating their reason, such that the participants can divulge whatever level of detail they deem appropriate). For the NM-reports, as these are given in response to an outcome the participant did *not* choose, it is altogether unclear what the basis of the report is, and if indeed we should predict that the participant would have anything at all to say.

However, we found no significant differences with respect to absolute word count. Another way to measure specificity is to count the number of unique words (that is, words only used once, in total 761 in the corpus). This division cuts through all word classes as a measure of relative rarity. But no significant differences between the NM- and M-reports were found on this measure either.

An alternative and more complex measure of the specificity of the statements is to look at the entire report, and determine to what extent the participants actually are talking about the choice they have made, and how much they are just (plain) talking. Following the guidelines of Brown and Yule (1983) we cleaned the corpus from all parts of the reports that did not involve a chain of reasoning, or listing of details that the participants thought had influenced their choice, thus separating the text into *content* and *metalingual comments*. Overall, around 50% of all transcribed text was classified as not strictly being about the choice, but this number did not differ significantly between the NM- and M-reports. Thus, the participants seemed to have as much content to report on regardless on whether they talked about a choice they had actually made, or responded to a mismatched outcome in a choice blindness trial.

Yet another way to get a grip on potential differences in specificity is to focus only on the amount of nouns used. This class of words contains all the details and features that surface in the participants' descriptions, such as "the face," "the eyes," "the hair." For the short reports we found no differences, but for the long reports there was a significant difference (Mann–Whitney $U = 3859$, $p = .019 < .05$) between the NM- and M-reports. The direction of the difference was also in line with the initial hypothesis—i.e., the relative frequency of nouns was higher in the NM-reports (mean = 0.089) than in the M-reports (mean = 0.078).

[5] We also calculated this contrast using a more inclusive set of words related to uncertainty, but no significant effects could be found with this measure either (see Table 2).

This is an interesting finding that raises the question of whether the dimension of specificity can also be discerned *within* the class of nouns, or if it lies more in the use of nouns as such. To investigate this, we listed all nouns from the material, and let two independent raters divide them into two groups.[6] One category concerned *specific* nouns, with words describing detailed features of the presented faces, such as *eyebrows* (ögonbryn), *haircut* (frisyr), *earrings* (örhängen), and *smile* (leende). The other category contained more *general* nouns, like *face* (ansikte), *picture* (bilden), *girl* (tjej), and *shape* (form). We tested these two categories separately, for both the short and the long reports, but with this measurement we found no significant differences for any of the conditions or categories.[7]

As a final test for specificity, we examined the generality of the noun difference, by running the same kind of analysis on the corpus of verbal reports collected in the Johansson et al. (2005) study. Using the current analysis as a template, we created a corresponding list of nouns for that material, divided into specific and general nouns (again, using two independent raters). Here, we found no significant differences between the NM- and M-reports, neither for nouns as a word class, nor for the division between specific and general nouns.

In summary, as in Johansson et al. (2005), we could not find any significant differences on the gross features of specificity for the NM- and M-reports, but for the more precise measurement of number of nouns used, a significant difference could be found for the long reports only (however, this difference could not be pinpointed to the use of more specific nouns, and it did not generalize to our previous corpus of reports).

4.3. Emotionality

The level of emotional engagement (whether positive or negative) is another of the obvious candidates for analysis that we investigated in Johansson et al. (2005). It is an obvious dimension to investigate because it is supposed to be present in the task (i.e., we would simply not have been so keen to compare the NM- and M-reports if it concerned a choice that the participants believed to be pointless). It is also a dimension that ought to be resistant to the manipulation, because even if the original reasons and intentions of the participants might be lost in the murky depths of their minds, at least they ought to still *prefer* the face they originally chose, and thereby show a more positive attitude toward the images in the NM-trials.

When looking for differences in emotionality, we proceeded in a similar fashion as we did with specificity. First, we measured the amount of adjectives, having identified them as the word class with most relevance for the levels of emotional engagement that the participants displayed in their reports. For this overall measurement, we found no significant differences between the NM- and M-reports. Then, using two independent raters, we created two subdivisions of adjectives: positive words—*beautiful* (vacker), *happy* (glad), *cute* (söt)—and negative words—*tired* (trött), *boring* (tråkig), *sad* (sorgsen). For the negative adjectives we found no significant differences, but for the positive ones we found a significant difference for the long reports only (Mann–Whitney $U = 3837.5$, $p = .0164 < .05$), such that there were more positive adjectives in the NM-reports (with the mean = 0.0474 for NM-reports, and the mean = 0.0367 for the M-reports). As with the previous finding for nouns, this difference did not generalize to the corpus collected in Johansson et al. (2005).

As we discussed above, this is a difference that makes a lot of sense in terms of the situation. Participants ought to show a more positive attitude toward the face they actually chose. But as emotionality is such a salient feature of the choice situation, both at the time of the original deliberation and at the time when the verbal report is given, this finding is not the best option for a clean indicator of the distinction between truthful and confabulatory report. This is so because for the full minute of speech delivered in the long reports, there is ample time for the original preference to assert itself, and for the participants in both the NM- and M-trials to *add* features to their report (while this concerns only minute differences, on average the NM-trials ought to build up in a more positive direction than the M-trials would).

[6] The interrater reliability for this task was very high, and for the few instances where the raters differed in their opinion, the disagreement was solved through further discussion among the raters. A similar procedure was used for all instances of independent rating mentioned in this article.

[7] If we glean at the mean value, we can see that there are 'unsignificantly' more specific *and* non-specific nouns in the long NM reports; a difference that in combination creates the overall significant difference for nouns. So the difference does not consist in the NM reports being more specific per se, just that more descriptive nouns in general are used.

In summary, we found a significant difference in positive emotional adjectives used between the NM- and M-reports for the long condition only. However, this difference is of unclear origin, and we could not replicate the finding in the corpus used in our earlier study.

4.4. Deceit

One line of inquiry that could potentially be of great use in contrasting and understanding the NM- and M-reports is research on the linguistic markers of deceit and lying. Even though the (possibly) confabulatory reports given by the participants in the M-trials obviously cannot be equated with an act of conscious and deliberate lying, it could be argued that the two situations share many features; most importantly, that something with no grounding in actual experience is being talked about.

The idea that statements derived from memory of an actual experience differ in content and quality from statements based on invention or fantasy has been the basis for several different methods for detecting deceit, such as criteria-based content analysis (CBCA, originally developed as a technique to determine the credibility of children's witness testimonials, Steller & Köhnken, 1989), and Reality Monitoring (RM, originally a paradigm for studying false memory characteristics, see Johnson & Raye, 1981). More recently, with the advent of powerful computers for large-scale data mining, this concept has blossomed into a separate field of automated deception detection (for overview, see Zhou et al., 2004a).

As an example of this development, Newman, Pennebaker, Berry, and Richards (2003) used Pennebaker's LIWC to distinguish between lies and truthful reports. In one of the conditions in this study, the participants were instructed to provide true and false descriptions of people they really liked or disliked. The deceptive element was thus to describe a person they really liked as if their feeling was very negative (and similarly, in the opposite direction for someone they disliked). Across all conditions, the software detected several persistent features that reliably predicted which statements were true and which were false. The variables they found to be primarily responsible for the differentiation were that liars used fewer first person references, fewer third person pronouns, fewer exclusive words (''except,'' ''but,'' ''without''), and more negative emotion words.

We were able to look directly at several of the critical variables identified by Newman et al. (2003). In particular, as there ought to be no real sense of ''me'' having preferred the outcome presented to the participants in the M-trials, we deemed the ''cognitive distance'' effect for first person references to be a good candidate to be represented in our material (what also has been called *verbal immediacy*, see Zhou, Burgoon, Nunamaker, Jay, & Twitchell, 2004b). We indexed all first person pronouns *I* (jag), *me* (mig), *my* (min/mina/mitt), *mine* (min/mina/mitt) in the corpus. These words were highly frequent, with *I* being the most frequent of all (with 1406 instances in total). We also counted all third person pronouns as an index of third person references (dominated by *she/her* (hon, henne), but also including *it* (den, det), *they* (dom) and *her* (hennes). In our corpus, we were unable to find an equivalent to the ''exclusive words'' category used by Newman et al. (2003).

However, despite verbal immediacy being a reliable predictor of deception, we found no significant differences for first person vs. third person pronouns between the NM- and M-reports (or for the negatively toned adjectives, as reported in the previous section on emotionality).

In summary, we found no significant differences between the NM- and M-reports by measuring them against linguistic markers of deceit.

4.5. Complexity

Another more theoretically driven perspective on the potential for the detection of markers of deceit in linguistic corpora is the assumption that lying is a more cognitively taxing activity than truthful report. Here, what is normally seen as markers of deceit should rather be seen as markers of *cognitive load* (Vrij, Fisher, Mann, & Leal, 2006). Evidence for this position comes from the fact that when training interrogators to detect deceit, it is more effective to instruct them to look for signs of the subjects ''thinking hard,'' rather than signs that they seem nervous or emotional (Vrij, 2004). But theories of cognitive load are obviously not confined to the field of deceit detection. It is one of the most widespread and most commonly used concepts in the cognitive sciences (and central to the whole idea of consciousness as a limited channel process, see Baars, 1997; Dehaene & Naccache, 2001). Translated to the task of introspective reporting in our choice-blindness paradigm, it lies close at hand to hypothesize that the participants in the M-trial would show a marked reduction in the *complexity* of the language used, as their resources ought to be taxed to a greater degree by the demands of reporting the reasons behind a choice they did not in fact make. For example, Butler et al., 2003) have reported a result close to this when showing that participants tend to use less complex language in a conversation task when they are simultaneously required to suppress a negative emotion.

The first and most simple way of measuring the complexity of NM- and M-reports is to look at the word length (e.g. Zhou et al., 2004b), where longer words are believed to require more effort to use. We calculated the mean word length for each of the four conditions, but we found no significant differences on this measure (short mean NM = 4.3 $M = 4.4$, long mean NM = 5.2, $M = 5.3$).

Two more advanced approaches to sentence complexity are the sibling concepts of *lexical density* and *lexical diversity*. What is meant by lexical density is essentially how informationally "compact" a text is (measured as the number of content words in relation to the number of grammatical or function words, Halliday, 1985; Ure & Ellis, 1977).[8] Lexical diversity, on the other hand, captures the uniqueness of the words used, i.e., how many different words there are in relation to the totality of the text (Malvern, Richards, Chipere, & Durán, 2004).

In our corpus we measured lexical density as the percentage of content words (nouns, verbs, adjectives, and adverbs) to all the words in a given text (content words plus grammatical words). Based on the hypothesized increase in cognitive load in the M-reports, it follows that they ought to have a lower lexical density. As we had already found differences in the base frequency of nouns and (positive) adjectives, it seemed as if this measure was a good candidate to reveal differences on a more structural level as well. However, we found no significant differences in lexical density between the NM- and M-reports.[9]

To measure lexical diversity we used the D algorithm from the CLAN software suite.[10] The sampling procedure used when calculating the measure D needs a minimum of 50 words for each entry. Given this constraint, we were only able to determine the lexical diversity for the long reports. But as was the case with lexical density, we found no significant differences between NM- and M-reports for this measure.

One interesting possibility here is that potential differences between the NM- and M-reports on lexical diversity are masked by a *priming effect*, such that novel words introduced during the NM-trials remain in an active state, and carry over to the (supposedly content-free) M-trials (i.e., this would be another way of stating the hypothesis that the cognitive load of the M-trials would reduce the complexity of the language used). We investigated this hypothesis by looking at the order in which the verbal reports were given for each participant, and calculating the number of new nouns introduced relative to what the participants had said before. However, the number of new nouns introduced did not significantly differ between the two conditions.

A final approach to unraveling the complexity of the introspective reports given by our participants would be to look at the *tense* and *themes* (i.e., structures of reasoning) they use to describe the chosen picture. There is no uniform way in which the participants use tense when explaining the reasons for the choices they have made. Sometimes they speak in the present tense, focusing on details in the preferred face ("she has such a round little nose"). But they can also refer back to the time of decision ("I liked her eyes and mouth"), or use comparative statements, in both past and present tense ("she had darker hair and she has so clear and pretty eyes"). The reasoning behind this measurement is again based on the concept of cognitive load. With less resources to spare in the M-trials, features of the current situation ought to have a greater impact on the report given (this could also be stated more intuitively as the idea that participants ought to speak more in

[8] A standard example of differing lexical density is between written and spoken text, in which written text normally has a larger proportion of content words (Halliday, 1985).

[9] It is interesting to note that there were differences between the *short* and the *long* reports, with the short reports being significantly more dense ($p = .007$).

[10] Intuitively, we can sense that there is a difference between for example the lush and varied style of Isabel Allende, and the stern and compact prose of Hemingway. But how to best capture such differences quantitatively is somewhat disputed (Malvern et al., 2004). The standard way of measuring diversity is type/token ratio (TTR) (i.e., the sentence "I am what I am" has three types and five tokens). However, as is now known, this method has certain statistical weaknesses. The best current alternative is the measure D, which we use here (Durán et al., 2004). So far, D has mainly been used to study language development, but it has also been put to some use in comparative studies on specific language impairment (SLI) and second language acquisition (Malvern et al., 2004).

present tense in the M-reports because they have no reason to refer back to from the moment the decision was made).

To investigate tense and themes we first created a basic index of all words related to tense (is/was, has/had, etc.), but we found no differences between the NM- and M-reports using this measurement. Next, to get a more precise measurement, we used the division between content parts and metalingual comments discussed in Section 4.2 above, and indexed the content part of the reports into either *positive reasons* for choosing the way they did, or *comparative reasons* why they preferred one face over the other one. Then these two categories were in turn divided into past and present tense.[11] But again, we found no significant differences between the NM- and M-reports.

In summary, using the concept of cognitive load and language complexity, we were unable to find any significant differences between the NM- and M-reports.

5. Latent semantic analysis

The differences we have found so far between the NM- and M-reports, using a whole battery of potential linguistic markers identified from the literature, have been small and very hard to interpret. But it is easy to envision that our search has been overly constrained by a limited theoretical outlook, or that is has been hampered because we lack crucial knowledge about some aspects of the relevant field of linguistics. Also, it could be argued that the ‘‘atomic’’ approach of word-frequency analysis is ill suited to capture differences of a more abstract semantic nature.

To allay these worries we decided to approach the corpus using a complementary bottom-up approach. Recent advances in computational cognitive analysis have opened up the intriguing possibility of quantifying semantics by applying advanced statistical techniques to huge text corpuses. These techniques are based on the postulate that semantics is carried by co-occurrences—that is, if two words frequently occur together in the same context (e.g., *love*-*like*), then this will be taken as evidence that the words have a similar meaning, or lie near each other in the semantic space.

Semantic spaces that include the semantic relationships of words from an entire language can be constructed using a method called *Latent Semantic Analysis* (LSA) (Landauer & Dumais, 1997). The way LSA works is that first a table for co-occurrence is created, where rows represent unique words and columns represent the contexts (e.g., sentences, paragraphs, or documents) from which the words are taken. Words that co-occur in the same context are marked with their frequency, otherwise a zero is marked. This table is then rescaled to account for differences in frequency by the logarithm of the frequency, and by dividing by the entropy across context. Finally, a semantic space is constructed by applying a mathematical technique called singular value decomposition (SVD) to reduce the large number of contexts to a moderate number of dimensions, all the while maintaining the maximal possible amount of the original information. The dimensions obtained correspond to the psychological concept of features that describe semantic entities in the words. The quality of the resulting semantic space can then be verified by applying a synonym test (and this information can in turn be used to further optimize the technique after optimization the number of dimensions left is typically found to be in the order of a few hundred, see, e.g., Landauer & Dumais, 1997).

Semantic spaces have successfully been applied in a number of linguistic and memory settings. Semantic spaces based on LSA have been shown to perform comparably to students in multiple-choice vocabulary tests, and in textbook final exams (Landauer, Foltz, & Laham, 1998). By measuring coherence, semantic spaces have also been used to predict human comprehension equally well as sophisticated psycholinguistic analysis (Landauer, Laham, & Foltz, 2003). In the domain of information search, LSA has also been found to improve retrieval by 10–30% compared to standard retrieval measure techniques (Dumais, 1994). Similarly, LSA has been used successfully to differentiate documents. As an example, Landauer, Laham, and Derr (2004) used

[11] As it is very hard to divide spoken text into discrete chunks (it is a close to arbitrary decision to decide where one statement ends and the next starts), we did not count the relative number of statements in past or present tense, but only measured whether it occurred or not in each verbal report. This means that the mean values presented in Table 2 are to be understood as the number of reports in which *some* parts were in past or present tense (and Why- or Comparative statements), which also means that the same verbal reports can feature in all of the four conditions at the same time. For this comparison between the NM- and M-reports, χ^2 was used as the statistical method.

sophisticated projection techniques to visualize scientific articles from different fields by projecting the high-dimensional semantic space to two-dimensional maps.

Taken together, these results indicate that LSA is an extremely promising tool for analyzing the semantic aspects of texts. However, currently there are no methods available for quantitatively comparing the semantics of two different classes of verbal report data, and for visualizing the results in a clear and convincing manner. Here, we introduce a new implementation of LSA specifically developed for this purpose, and apply it to the corpus of reports collected in the choice-blindness paradigm.

5.1. Method

As a base corpus, the Stockholm-Umeå Corpus (SUC, Ejerhed & Källgren, 1997) consisting of one million Swedish words was selected. This corpus is balanced according to genre, following the principles used in the Brown and LOB corpora. Infomap (http://infomap.stanford.edu/), a natural language software that implements LSA, was then used to create a semantic space. Context was defined as 15 words before, or after, the current word in the present document. Following initial testing, we settled for a space consisting of 150 dimensions. The length of the vector describing each word was normalized to one.

The semantic spaces were processed in LSALAB,[12] a program specifically developed by one of authors to analyze semantic spaces. Each verbal justification for choosing a particular face was summarized to one point in the semantic space by averaging the semantic location of all the words included in the statement. To be sure that the semantic representations were stable and reliable, we included only the 4152 most common words from the SUC corpus (words with lower frequency were ignored).

As we are unaware of any other studies applying statistical methods to compare conditions within a semantic space, we developed the following technique to handle the issue. The semantic point describing each condition (e.g., NM- and M-trials) was summarized as the average of the semantic points of all statements included in the condition. The Euclidean distance was then used as a measure of distance between the conditions (μ_1). After this, a bootstrap technique was applied to estimate the variability in distance. Statements were randomly placed in either of the two conditions (using the same number of trials), and the distance was calculated. To achieve a reliable estimate this was repeated for 200 trials. A one-tailed t-test was calculated by subtracting the mean distance of the random trials (μ_0) from the distance between the conditions (μ_1), and this was then divided by the estimated standard deviation of distance for the random trials (σ).

As LSA deals with a multi-dimensional space, graphic illustration is essential to understanding the results. However, the plotting of such high-dimensionality spaces is problematic, as it typically requires a projection to only two dimensions.[13] To deal with this problem we propose the use of a two-dimensional separation–typicality map. These maps are obtained by the following method.

We base both of the axes on the Euclidean distance, where the x-axis represent *separation* and the y-axis *typicality*. Separation on the x-axis is based on a distance measure that maximally differentiates between the two conditions. The natural choice is the distance from a statement to the prototype of one of the conditions. To separate condition 1 and 2, we simply plot the difference in distance (DID), which is the Euclidean distance from a statement to the prototype of condition 1 minus the Euclidean distance from same statement to the prototype of condition 2. However, the DID measure is subject to a statistical artifact. Because the instances are compared with the prototype, the separation between the conditions will be inflated. This artifact

[12] For details, see http://www.lucs.lu.se/People/Sverker.Sikstrom/LSALAB_intro.html.

[13] Landauer et al. (2004) argue for the visualization of semantic spaces as a powerful tool for understanding, viewing, and exploring semantic data. They were able to plot the semantic representation of more than 16,000 scientific articles from *Proceedings of the National Academy of Sciences* (PNAS) using the GGobi software (Swayne, Cook, & Buja, 1998). In this case, dimensionality reduction was conducted by a combination of mathematical tools and visual inspection. Although this procedure was successful in separating and finding sub-cluster in the data space, it has several problematic aspects to it. First, the choice of a projection to a low-dimensional space can be made in an almost infinite number of ways, so the resulting conclusion becomes highly dependent on this choice. Second, while choosing projections, statistical artifacts may bias the separation between conditions so they appear to be larger than they actually are. For example, separating two conditions sampled from the same population for 100 dimensions will results in an expected value of 5 statistically different dimensions due to chance. Plotting these dimensions will amount to a form of data fishing, and the separations will only be statistical artifacts. Third, when using the Landauer et al. (2004) methodology, the axes on the plot are not immediately available for interpretation.

can be removed by a bootstrapping technique whereby the statements are randomly placed into the two conditions. To obtain sufficient statistics we repeated this 200 times. We then subtracted the average DID obtained from the random samplings from the DID of each statement. The resulting corrected DID value, which we label DID', is free from statistical artifacts, so that the expected value of the separation from randomly generated populations is zero. DID' is a measure of the separation between the conditions. If the two prototypes are identical then the value will always be zero.

On the y-axis we plot the typicality of the statements. This is simply the Euclidean distance between the statement and the prototype of all statements. This measure is bounded between zero and two in our semantic representation. A zero value indicates that the statement is identical to the prototype of all statements. A value of two indicates that the statement is maximally different from the prototype. A value of one indicates that the statement is unrelated to prototype (i.e., the expected value of a randomly generated statement). Most often the values will fall in the range 0 to 1, where low values indicate statements that are typical and high values indicate semantically atypical statements.

5.2. Results

There was no statistical difference in semantic content between the NM- and M-reports ($t(388) = -.91$; $p = .82 > .05$). Thus, the result of the statistical analysis of the semantic space indicates that the participants justify their choice using the same semantic content for both the NM- and the M-trials.

To visualize these results we use the separation–typicality map described above. Fig. 1 plots the separation between the statements on the x-axis, and the typicality (low values indicate high typicality) on the y-axis. Each dot represents a NM-report, and each cross an M-report. The large dot and cross represent the average values over all statements in each condition. The curves in the lower part of the graph are the densities of the respective condition. As is apparent from Fig. 1, the overlap between the NM- and M-reports is almost complete. The typicality of statements ranges from approximately 0.35 (high typicality) to 1.2 (low typicality), with a mean around 0.6, where 1 represents statements that are unrelated to the prototype of all statements.

While LSA is a well-established and powerful technique for building semantic spaces, it has never before been used for significance testing in this type of contrastive methodology. Thus, a possible reason for the lack

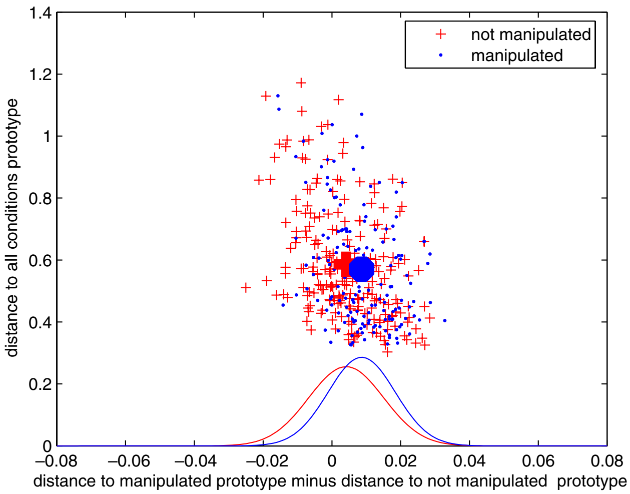

Fig. 1. Separation–typicality map for the NM- and M-reports. The y-axis plots the semantic distance to the prototype of all conditions as a function of difference in distance (DID) on the x-axis. Each cross and dot represent a manipulated or not manipulated statement, respectively. The large dot and cross represent the average values over all statements in each condition. The expected distance between two randomly semantic locations is one, and the maximally possible distance is two, compared with the distance to all conditions prototype on the y-axis. The difference in distance between the conditions on the axis represents the difference between the conditions, so that if the two conditions' prototypes were identical then the distance would be zero. The two curves in the lower part of the graph show the density of statements for the two conditions.

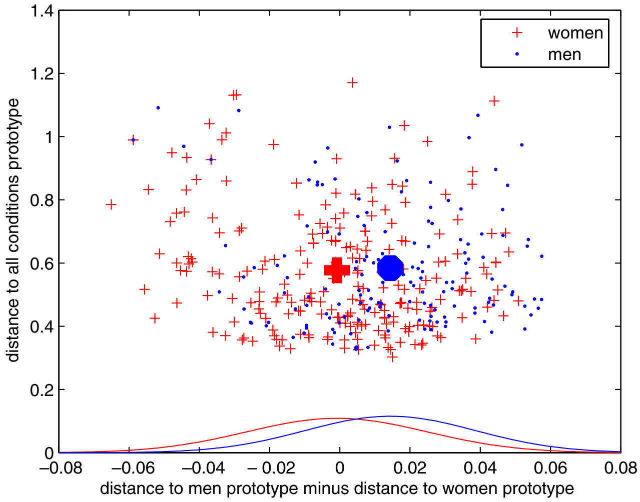

Fig. 2. Separation–typicality map for the female and male reports. Each cross and dot represent a male or female statement respectively. In all other regards, the figure is the same as Fig. 1.

of separation between NM- and M-reports could be that our proposed method is not sensitive enough to differentiate between the two conditions. In order to minimize this risk, it is important to demonstrate that the method indeed can detect meaningful differences under conditions where those differences are likely to emerge. To demonstrate this we ran the same kind of differentiation analysis using the gender of the participants as an input variable.[14] In contrast to what was the case for the NM- and M-reports, we found a highly significant difference between the introspective reports given by men and women ($t(388) = 2.98$; $p = .002 < .05$). Thus, it can be shown that the method we used is sufficiently sensitive to distinguish between the semantic content of statements produced by two contrast groups.

Fig. 2 shows a separation–typicality map for female and male reports. As is evident from the figure, a clear separation between the two groups of report can be found. This is reflected in the large variability on the *x*-axis (compare with the low variability in Fig. 1 showing the NM- and M-reports).

However, given that the statements made by men and woman differ in their semantic content, the question remains how best to characterize these differences. To try to capture the differences found, we listed all the words in the constructed semantic space that had the closest semantic location to the male and female prototypes, respectively. These associates may be conceived of as a type of "keywords" that summarize something about all statements in the conditions. The first thing to notice is that the keywords for statements made by men and women are highly similar (e.g., see the first two columns in Table 3). The first seven associates are identical (with the exception of a single flip of the ordering). This demonstrates that the similarities between the male and female reports are great, yet we are still able to discern the subtle differences residing in the material. This point further strengthens the inference that had there been any semantic content differences between the NM- and M-reports, it is highly likely that our method would have picked them up.

As explained above, one of the virtues of LSA is that it embodies very few assumptions about the nature of the subject under study. In this way, there is a greatly diminished risk that the results are contaminated by either common-sense intuitions, or the particular theoretical outlook of the experimenters. To identify more clearly the difference in the reports made by males and females, we subtracted the male and female prototype vector from each other. The closest semantic associates to this vector are listed in column four in Table 3. For females, out of the approximately four thousand possible words in our semantic space, the two highest associates were the female pronouns *her* and *she*. A large proportion of the remaining associates were body parts (*face*, *foot*, *hand*, *mouth*, *arm*). For males, the closest associates to this vector are shown in column three in

[14] As the experiment collected very few personality variables, the age-spread of the participating student population was limited, and each image-pair contained too few reports to be entered into the analysis, gender emerged as the best candidate variable to work with.

Table 3
The closest semantic associates to male and female prototypes

Associates		Differences	
Men	Women	Men	Women
It	It	Analysis	Hers
But	But	Interested	She
Not	Not	True	Face
I	Be	Democratic	Foot
Be	I	Doubt	And
To	To	Name	Down
Just	Just	Know	Fine
Have	Only	Pull	Hand
As	Have	Think	Out
Know	She	It	Skirt
Him	Accomplish	Hardly	Mouth
What	He	Starting-point	Kiss
Become	And	What	Sit
And	Become	Up	Arm

Note. The first two columns show the fourteen closest semantic associates to statements made by men and women respectively, starting with the closest associates. The last two columns show semantic associates to the vector describing the difference between the two prototypes, where the column labeled men is the closest associate to the vector men minus women, and the column labeled women the vector women minus men. It is important to stress that none of the words displayed in the columns actually needs to be represented in the choice-blindness corpus (i.e., no male participant need ever have used the word "democratic" when describing why they choose one face over the other). In this case the associates instead come from the million word SUC corpus used to anchor the semantic space. All words in the table are translated from Swedish to English.

Table 3. These associates tend to be more abstract (*analysis*, *democratic*), and revolve around the theme of knowing (*true*, *doubt*, *know*, *think*, *hardly*).

It is not possible to provide an exact summary of the semantic differences in associations between the gender specific statements, as there is no fully transparent mapping from the dimensions captured by LSA onto everyday concepts. But, as reported above, the outcome suggests a separation along a dimension of concreteness–abstractness, and into themes of knowing vs. body parts, and in the particular use of personal pronouns. However, these results are far from the end-point of the inquiry. They should rather be seen as a kind of *data-driven hypothesis generators*. For validation and translation into everyday concepts, additional work would be required that attempted to further quantify and test the identified dimensions.[15]

6. How something can be said about telling more than we can know

It probably has not escaped the reader that this article has an unusual format for the presentation of the main results—i.e., we treat the failure to find distinguishing markers between the NM- and M-reports as an equally important finding as any of the potential differences found. We are aware that, from a textbook perspective, this logic is clearly flawed (i.e., with standard significance testing, the null hypothesis cannot be confirmed, only rejected), yet we cannot escape the conclusion that the overall pattern of findings indicates that the NM- and M-reports are surprisingly similar. To really appreciate this null-hypothesis blasphemy, we must

[15] For example, if we compare these results to the more than twenty significant differences that we found between the male and female reports using the categories previously reported for the word-frequency analysis, the complementary, but also partially overlapping, character of the LSA analysis becomes obvious. Regarding the female LSA associates for the female pronouns, a match can be found with the word-frequency analysis that indicated a higher degree of use of personal pronouns by women (short reports, Mann–Whitney $U = 3447$, $p = .026 < .05$). The LSA differences between females and males for the dimension of concreteness–abstractness also seems to be reflected in the word-frequency analysis, where we found females to be using more specific nouns (long reports, Mann–Whitney $U = 3678.5$, $p = .004 < .05$), and more non-specific nouns (short reports, Mann–Whitney $U = 3379$, $p = .016 < .05$). However, the knowing-theme from the LSA analysis does not seem to have an immediate counterpart among the epistemic measures used in the word-frequency analysis, and there are also several other significant differences from the contrastive linguistic analysis that did not emerge in our global LSA comparison (i.e., word length, high-low frequency words, present tense, pauses, prepositions, conjunctions, etc.).

go back to the sentiments we had, and the predictions we made (including those of our colleagues) before we conducted our first choice blindness experiment. Tentatively stating a hypothesis at this time, we predicted not just differences between the NM- and M-reports, but *huge* differences. As it stands now, not a single difference found in the current corpus would survive a standard Bonferroni correction.[16] This can be compared to the strong pattern of differences between male and female reports, which we were able to discern both with word-frequency analysis and with LSA.

Another way of framing the subtlety of the possible differences between NM- and M-reports existing in our material is by comparing them to the literature on automatic lie detection we briefly referenced in Section 4.4. For detection of lies based on linguistic cues only, Newman et al. (2003) and others (e.g., Zhou et al., 2004b), have shown that prediction models can be built that capture general differences between truths and lies using very similar dimensions to those measured in this article (i.e., certainty, emotionality, complexity, etc.). It is a telling point that the differences in the deceit literature are so small that untrained human observers basically predict at chance level, while finely calibrated software only reaches levels of predictability of about 60–65% (Newman et al., 2003). However, for the contrast between the NM- and M-reports in our material it is at present doubtful whether *any* such model can be built.

We believe we have conducted a thorough and revealing investigation of the introspective reports collected so far in our choice blindness paradigm. Including the analysis done in Johansson et al. (2005), we have used three complementary types of measurement (psychological rating, word-frequency analysis, and LSA), and all three have come out with very similar results.

But obviously, this is just a starting point. For example, the fact that the two tentative differences we found in the material (on specificity and emotionality) only could be found for the long reports might suggest that one should look more closely at *time* as a factor in future studies. However, the remarkable thing from our perspective is that the debate about the nature and validity of introspection is still conducted at a level where the introduction of a contrast class between (potentially) genuine, and (potentially) confabulatory reports seemingly can tell us a great deal about what introspection amounts to. A simple contrastive methodology is often derided by researchers from more mature fields of science, but it can still function as a springboard for other more penetrating approaches (as has been the case with lesion studies, studies of individual differences, cross-cultural comparisons, etc.). In this sense, Nisbett and Wilson (1977) were far ahead of their times when they introduced a methodology that required the experimenters to *know and control* the causes of the behavior of the participants for it to work. N&W strove admirably for ecological validity in their experiments, but 30 years later (notwithstanding the wet dreams of some marketers and retailers) this is still something the behavioral sciences are incapable of doing, save in the most circumscribed and controlled environments.

In this vein, it can be seen that the most famous of the experiments of N&W, the department-store stocking experiment, involved a rather strange and contrived task (e.g., Kellogg, 1982; Kraut & Lewis, 1982). It seems to us, had only the experimenters had a better grasp of what influenced the choice behavior of normal consumers, they would not have given them the artificial choice between *identical* stockings, but rather something that would have involved actual products of varying quality.

While we do not want to pretend that the task we have used here (and in Johansson et al., 2005) involves an important choice for the participants, it is a very straightforward one, reflecting a type of judgment that people often make in their daily lives (and undoubtedly, many people have strong opinions about facial attractiveness). It has the virtue of being a simple and vivid manipulation that does not place the same exorbitant demands on the experimenters to be able to secretly influence the decision process of the participants. Like the hypothetical "intuition pumps" so often employed in debates about consciousness and introspection (see Dennett, 1991), this is an experiment where it is child's play to twiddle with the knobs (parameters) of

[16] Bonferroni correction is a commonly adhered-to guideline when doing exploratory research, a safeguard to prevent results arising from chance fluctuations when multiple tests of statistical significance are done on the same data set. It states that for multiple comparisons the p level should be equal to alpha-level/number of observations ($0.05/N$). As more than 30 variables are measured in this article (for both short and long reports), even if not adhered to strictly, none of the seemingly significant results are firm enough to remain after a Bonferroni correction. The reason we did not include this calculation in the results section is that we prefer to err on the side of including non-existent differences, rather than the other way around. As this type of contrast has not been made before, we believe it to be of great importance to grasp every straw there is to generate further hypotheses about how the NM- and M-reports might relate to each other.

the setup, and produce potentially very interesting results (by changing properties of the stimuli, deliberation time, questions asked, context of choice, personality variables, etc.).

Philosophically speaking, our choice blindness paradigm is of the same breed as the N&W experiments. We believe it to be an improvement over N&W in many regards, but at this point there are many opportunities for interpretations open for the wily theoretician. For example, the fact that we can hardly find any differences between the NM- and M-reports could stem from the participants actually reporting the very same thing in both conditions—i.e., the intentions they had for making their actual choice. But this is a strained interpretation to make when one sees how good the match between the given reports and the presented faces often are, and it creates outright absurdities in those cases where the reports refer to unique features of the manipulated face (e.g., "I chose her because I love blondes," when in fact the dark-haired one was the chosen one). Conversely, when differences between NM- and M-reports are found, they could have been created at the time of actual reporting, rather than being inherited from the deliberation phase. As we discussed briefly in the section on emotionality, the interaction of prior preferences and the outcome of the choice could possibly lead the two classes of reports to diverge (i.e., in the M-trials the participants are reacting to a face they did not prefer, no wonder then they are not exuberant about it now).

It is also clear that the simplification we have made in this article, where we keep the analysis of the verbal reports more or less separate from the basic choice blindness effect, cannot be maintained in the long run. If we are to fully understand introspection, then we should be prepared to explain the whole architecture of a decision-making system in which one might fail to notice mismatches between intention and outcome, but yet give perfectly intelligible verbal reports in response to the manipulated choice. However, as we said in the introduction, we have an upbeat outlook on the prospects for development in this field. It seems to us that the simple contrast at the heart of our choice blindness paradigm is perfectly poised to be used in the kind of triangulation of subjective reports, behavioral responses, and brain imaging data that Roepstorff and Jack (2004) identify as the best route for future studies of introspection and consciousness to take.

In conclusion, we want to emphasize the potential of our method over the particularities of the results in this article. When Nisbett and Wilson (1977) took upon themselves not only to introduce a new experimental paradigm, but to formulate a theory of introspection in sharp contrast to the prevailing view, they set the research community up for a high-strung showdown, not unlike the archetypal movie scene where the protagonists suddenly find themselves locked at mutual gunpoint (the so-called "Mexican standoff"), and where the smallest twitch of the pen inevitably will release a hail of deadly arguments. In our minds, far too little has been said about telling more than we can know, for us to have reached a point where a standoff is called for. Instead, it is our hope that the effort put forward here will lead to a renewed interest in experimental approaches to the study of verbal report and introspection.[17]

Acknowledgment

We would like to thank Jordan Zlatev, Victoria Johansson, Joost van de Weijer and Mats Andrén for all their help and advice. The work of LH was funded by the Erik philip Sörensen Foundation.

References

Allwood, J. (1998). Some frequency based differences between spoken and written Swedish. Paper presented at the XVIth Scandinavian Conference of Linguistics, Department of Linguistics, University of Turku.

Baars, B. (1997). *In the theater of consciousness: The workspace of the mind*. Oxford: Oxford University Press.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Brewer, B. W., Linder, D. E., Vanraalte, J. L., & Vanraalte, N. S. (1991). Peak performance and the perils of retrospective introspection. *Journal of Sport and Exercise Psychology, 13*(3), 227–238.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge University Press.

[17] If we allow the visionary movie industry to lead our way, in contrast to the spaghetti westerns of the 70s, the B-movie thrillers of the 80s, and the bloody mayhem of Tarantino in the 90s, the recent movie Munich (2005), contains a scene with a friendly resolution of an incredibly tense Mexican standoff.

Butler, E. A., Egloff, B., Wilhelm, F. H., Smith, N. C., Erickson, E. A., & Gross, J. J. (2003). The social consequences of expressive suppression. *Emotion, 3*, 48–67.

Dehaene, S., & Naccache, L. (2001). Toward a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition, 79*, 1–37.

Dennett, D. (1991). *Consciousness explained*. Boston: Little: Brown & Company.

Dumais, S. T. (1994). Latent semantic indexing (LSI) and TREC-2. Paper presented at the Second Text Retrieval Conference (TREC2).

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*(2), 220–242.

Ejerhed, E., & Källgren, G. (1997). Stockholm Umeå corpus (Version 1.0): Department of Linguistics, Umeå.

Elliot, W., & Valenza, R. (to appear). Two tough nuts to crack: Did Shakespeare write the Shakespeare portions of Sir Thomas More and Edward III? In Shakespeare yearbook.

Frawley, W. (1992). *Linguistic semantics*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Gavanski, I., & Hoffman, C. (1986). Assessing influences on one's own judgments: is there greater accuracy for either subjectively important or objectively influential variables. *Social Psychology Quarterly, 49*(1), 33–44.

Glenn, P. (2003). *Laughter in interaction*. Cambridge: Cambridge University Press.

Goldman, A. I. (2004). Epistemology and the evidential status of introspective reports. *Journal of Consciousness Studies, 11*(7–8), 1–16.

Guerin, B., & Innes, J. M. (1981). Awareness of cognitive-processes—replications and revisions. *Journal of General Psychology, 104*(2), 173–189.

Hall, L., Johansson, P., Tärning, B., Sikström, S. (in preparation). Choice Blindness and Preference Change: Lund University Cognitive Science.

Halliday, M. A. K. (1985). *Spoken and written language*. Deakin: Deakin University Press.

Higuchi, K. A. S., & Donald, J. G. (2002). Thinking processes used by nurses in clinical decision making. *Journal of Nursing Education, 41*(4), 145–153.

Hirstein, W. (2005). *Brain fiction: Self-deception and the riddle of confabulation*. Cambridge, Massachusetts: The MIT Press.

Holmes, J. (1995). *Women, men and politeness*. London: Longman.

Holmes, J. (1997). Women, language and identity. *Journal of Sociolinguistics, 1*(2), 195–223.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310*(5745), 116–119.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*, 67–85.

Jopling, D. A. (2001). Placebo insight: the rationality of insight-oriented psychotherapy. *Journal of Clinical Psychology, 57*(1), 19–36.

Jorgensen, A. H. (1990). Thinking-aloud in user interface design: a method promoting cognitive ergonomics. *Ergonomics, 33*(4), 501–507.

Kellogg, R. T. (1982). When can we introspect accurately about mental processes. *Memory & Cognition, 10*(2), 141–144.

Kraut, R. E., & Lewis, S. H. (1982). Person perception and self-awareness knowledge of influences on ones own judgments. *Journal of Personality and Social Psychology, 42*(3), 448–460.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Newbury Park, CA: Sage.

Labov, W. (1972). *Sociolinguistic patterns*. Oxford: Blackwell.

Lakoff, R. (1975). *Language and woman's place*. New York: Harper Colophon Books.

Landauer, K., Laham, D., & Derr, M. (2004). From paragraph to graph: latent semantic analysis for information visualization. *Proceeding of the National Academic of Science, 101*, 5214–5219.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Landauer, T. K., Foltz, P., & Laham, D. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*, 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. New York: Palgrave Macmillan.

Morris, P. E. (1981). The cognitive psychology of self-reports. In C. Antaki (Ed.), *The psychology of ordinary explanations of social behaviour*. London: Academic Press.

Nagel, T. (1974). What is it like to be a bat?. *The Philosophical Review 83*(4), 435–450.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665–675.

Nisbett, R. E., & Bellows, N. (1977). Verbal reports about causal influences on social judgments: private access versus public theories. *Journal of Personality and Social Psychology, 35*(9), 613–624.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review, 84*(3), 231–259.

Norrby, C. (2004). *Så gör vi när vi pratar med varandra* (2nd ed.). Lund: Studentlitteratur.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC2001*. Lawrence Erlbaum.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves. *Annual Review of Psychology, 54*, 547–577.

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: language use across the lifespan. *Journal of Personality and Social Psychology, 85*(2), 291–301.

Quattrone, G. A. (1985). On the congruity between internal states and action. *Psychological Bulletin, 98*(1), 3–40.

Rakover, S. S. (1983). Hypothesizing from introspections—a model for the role of mental entities in psychological explanation. *Journal for the Theory of Social Behaviour, 13*(2), 211–230.

Roepstorff, A., & Jack, A. I. (2004). Trust or interaction? Editorial introduction. *Journal of Consciousness Studies, 11*(7–8), V–XXII.

Rorty, R. (1993). Holism, intrinsicality, and the ambition of transcendence. In B. Dahlbom (Ed.), *Dennett and his critics: Demystifying mind* (pp. 184–202). Cambridge, Mass: Basil Blackwell.

Sabini, J., & Silver, M. (1981). Introspection and causal accounts. *Journal of Personality and Social Psychology, 40*(1), 171–179.

Sandberg, J. (2005). The influence of network mortality experience on nonnumeric response concerning expected family size: evidence from a Nepalese mountain village. *Demography, 42*(4), 737–756.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22*(2), 127–160.

Sprangers, M., Vandenbrink, W., Vanheerden, J., & Hoogstraten, J. (1987). A constructive replication of white alleged refutation of Nisbett and Wilson and of Bem: limitations on verbal reports of internal events. *Journal of Experimental Social Psychology, 23*(4), 302–310.

Steller, M., & Köhnken, G. (1989). Criteria-based content analysis. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 217–245). New York: Springer-Verlag.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and non-suicidal poets. *Psychosomatic Medicine, 63*, 517–522.

Stone, L. D., & Pennebaker, J. W. (2002). Trauma in real time: talking and avoiding online conversations about the death of Princess Diana. *Basic and Applied Social Psychology, 24*, 172–182.

Swayne, D. F., Cook, D., & Buja, A. (1998). Xgobi: interactive dynamic data visualization in the X window system. *Journal of Computational and Graphical Statistics, 7*, 113–130.

Ure, J., & Ellis, J. (Eds.). (1977). *Register in descriptive linguistics and linguistic sociology*. The Hague: Mouton Publishers.

Vartatala, T. (2001). Hedging in the scientifically oriented discourse. Exploring variation according to discipline and intended audience. Ph. D. Thesis.

Vold, E. T. (2006). Epistemic modality markers in research articles: a cross-linguistic and cross-disciplinary study. *International Journal of Applied Linguistics, 16*(1), 61–87.

Vrij, A. (2004). Why professionals fail to catch liars and how they can improve. *Legal and Criminological Psychology, 9*, 159–183.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences, 10*(4), 141–142.

White, P. A. (1987). Causal report accuracy: retrospect and prospect. *Journal of Experimental Social Psychology, 23*(4), 311–315.

White, P. A. (1988). Knowing more about what we can tell: introspective access and causal report accuracy 10 years later. *British Journal of Psychology, 79*, 13–45.

Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. London: The Belknapp Press.

Wilson, T. D., & Kraft, D. (1993). Why do I love thee—effects of repeated introspections about a dating relationship on attitudes toward the relationship. *Personality and Social Psychology Bulletin, 19*(4), 409–418.

Wilson, T. D., Laser, P. S., & Stone, J. I. (1982). Judging the predictors of ones own mood: accuracy and the use of shared theories. *Journal of Experimental Social Psychology, 18*(6), 537–556.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*(1), 101–126.

Wilson, T. D., & Stone, J.I. (1985). Limitations of self-knowledge: More on telling more than we can know. In S.P. (Ed.), Review of personality and social psychology: Self, situations, and social behavior (Vol. 6, pp. 167–185).

Wright, P., & Rip, P. D. (1981). Retrospective reports on the causes of decisions. *Journal of Personality and Social Psychology, 40*(4), 601–614.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., Nunamaker, J., & Jay, F. (2004a). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems, 20*(4), 139–165.

Zhou, L., Burgoon, J. K., Nunamaker, J., Jay, F., & Twitchell, D. (2004b). Automatic linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation, 13*, 81–106.