Measuring Cues to Deception: A Multitrait-Multimethod Analysis

by

Ryan Lahay

A thesis submitted to the
School of Graduate and Postdoctoral Studies in partial
fulfillment of the requirements for the degree of

Master of Science in **Forensic Psychology**

Faculty of Social Science and Humanities

University of Ontario Institute of Technology (Ontario Tech University)

Oshawa, Ontario, Canada

December 2021

THESIS EXAMINATION INFORMATION

Submitted by: **Ryan Lahay**

Master of Science in **Forensic Psychology**

Thesis title: Measuring Cues to Deception: A Multitrait-Multimethod Analysis

An oral defense of this thesis took place on December 1st, 2021 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee	Dr. Leigh Harkins
Research Supervisor	Dr. Amy-May Leach
Examining Committee Member	Dr. Matthew Shane
Thesis Examiner	Dr. Bobby Stojanoski, *Ontario Tech University*

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

Abstract

Cognitive load and arousal are constructs typically included in theories of deception, but they are often measured using a range of unvalidated techniques. Using a multitrait-multimethod analysis, I assessed the reliability and construct validity of common measures of cognitive load and arousal – self-report, trained coders' observations, and behavioral measures – across three studies as secondary data. All measures showed good reliability, but achieved differing levels of validation. Measures of cognitive load (i.e., self-reported cognitive load, trained coders' observations of thinking hard, and average response latency) showed some evidence of construct validity. In contrast, measures of arousal (i.e., self-reported arousal, trained coders' observations of nervousness, and average skin conductance) did not achieve sufficiently high levels of validity. These findings suggest that researchers may not be assessing constructs of interest. Thus, researchers should exercise caution when using unvalidated measures to evaluate theories of, and the diagnosticity of cognitive and arousal-based cues to, deception.

Keywords: cues to deception; validity; multitrait-multimethod

Author's Declaration

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Ontario Institute of Technology (Ontario Tech University) to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Ontario Institute of Technology (Ontario Tech University) to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

Ryan Lahay

Statement of Contributions

This thesis has been submitted for publication:

Lahay, R., Leach, A-M., Cutler, B. L., Woolridge, L. R., & Elliott, E. (2021). *Measuring cues to deception: A multitrait-multimethod analysis*. Manuscript submitted for publication.

I contributed to the conceptualization of the research questions and research design, performed some of the data collection and all analyses, and wrote the manuscript. I have used standard referencing practices to acknowledge ideas, research techniques, or other materials that belong to others.

Acknowledgements

To my supervisor, Dr. Amy Leach, thank you for your persistent support, time, wisdom, guidance, encouragement, and optimism. I will forever be thankful for all that you have done to contribute to my success. You have been an exceptional supervisor and mentor, and I believe that I am a better researcher as a result.

To Dr. Brian Cutler, thank you for your insight and support in developing this project, as well as your mentorship throughout my graduate studies.

To Dr. Leigh Harkins, Dr. Matthew Shane, and Dr. Bobby Stojanoski, thank you for your feedback and guidance on this thesis.

To the Social Sciences and Humanities Research Council (SSHRC) and sponsors of the Ontario Graduate Scholarship (OGS; Government of Ontario and donors), thank you for your financial support.

To my fellow graduate students and lab colleagues, thank you for contributing to an enjoyable graduate school experience, alleviating some of the stress that accompanied it, and answering all of my questions. I look forward to our continued friendships.

To my research assistants, thank you for your patience and tireless efforts in completing any projects I sent your way. It is greatly appreciated.

To my girlfriend and our families, thank you for your unfaltering support, love, and patience during this journey.

Finally, I would like to thank anyone else who has supported me along the way and helped contribute to the development of this thesis and/or the person that I am today.

I appreciate each and every one of you.

Table of Contents

List of Tables

List of Figures

Measuring Cues to Deception: A Multitrait-Multimethod Analysis

Two constructs that are often posited to underlie deception are cognitive load and arousal (e.g., Zuckerman et al., 1981). Several cues to deception have been associated with each, although findings have been relatively heterogenous across individual studies (see DePaulo et al., 2003, but also Luke, 2019). Importantly, the measures that have been used to assess cognitive load and arousal have also been inconsistent. Self-report, observers' ratings, and behavioral techniques have been used relatively interchangeably (Vrij, 2008), yet their equivalence has not been tested empirically. The discrepancy in measures across studies may, thus, be a source of variability in cue diagnosticity and, more broadly, theory evaluation. As such, a critical step toward advancing theories of deception is to determine whether commonly used measures of cognitive load and arousal are equally valid.

Processes Underlying Deception

There is no unifying theory of deception detection. Indeed, several competing theories have specified the processes underlying deception and their associated cues. Elaborating upon each is beyond the scope of this paper; however, some of the most influential theories to date – including the Multifactor (or Four Factor) Model (Zuckerman et al., 1981), the Self-Presentational Perspective (DePaulo et al., 2003), Interpersonal Deception Theory (IDT; Buller & Burgoon, 1996), the Working Memory Model (Sporer, 2016; Sporer & Schwandt, 2006), and the Activation-Decision-Construction-Action Theory (ADCAT; Walczyk et al., 2014) – share key elements.

First, in each theory, cognitive load and/or arousal are central constructs. Regarding cognitive load, lie-tellers are consistently posited to experience greater

cognitive load than truth-tellers due to the demands associated with formulating a plausible lie (e.g., Sporer, 2016; Zuckerman et al., 1981). Although not all theories include arousal as an independent construct (e.g., Walczyk et al., 2014), the majority acknowledge the role of arousal in deception, even if only incidentally. In particular, lie-tellers are generally expected to experience greater physiological arousal than truth-tellers (e.g., DePaulo et al., 2003; Zuckerman et al., 1981). This arousal is further attributed to emotions associated with deception in some theories (e.g., Zuckerman et al., 1981).

Second, theorists have posited that cognitive load and/or arousal would give rise to measurable behavioral differences between lie- and truth-tellers. For example, according to the Multifactor Model (Zuckerman et al., 1981), lie-tellers' heightened cognitive load was predicted to increase delayed responses and speech hesitations. However, specific cues and the intensity and directionality of effects have varied between theories. Therefore, a critical component of theory validation should involve measuring these differences between lie- and truth-tellers.

Cues to Deception

The literature on the diagnosticity of cues to deception is extensive (see DePaulo et al., 2003; Sporer & Schwandt, 2006). No single behavior always signals deception. However, there have been cues that reliably distinguished between lie- and truth-tellers (DePaulo et al., 2003). Because cognitive load and arousal cues were of greatest relevance to established theories, and to my research questions, I focus solely on those here. Importantly, however, the classification of specific behaviors as cognitive- or arousal-based has not been universal and somewhat contradictory. For example, some researchers have argued that an increase in blinking may represent heightened arousal,

but a decrease in blinking may represent greater cognitive load (see DePaulo et al., 2003 for a discussion). Cues have been classified based on their most common usage in the field, particularly as established by Zuckerman et al.'s (1981) and DePaulo et al.'s (2003) seminal meta-analyses.

If theories positing that deception heightens cognitive demands (e.g., working memory; Sporer & Schwandt, 2006) are correct, then cognitive cues to deception should differentiate between lie-tellers and truth-tellers. Indeed, in DePaulo et al.'s (2003) meta-analysis, lie-tellers exhibited significantly fewer movements, claimed a lack of memory more often, sounded less interested / involved when speaking, and seemed more uncertain than truth-tellers. Lie-tellers also took significantly longer to respond to questions and exhibited more silent pauses than truth-tellers, but only when they could not prepare their responses or were questioned over a longer period of time. Sporer and Schwandt's meta-analysis replicated DePaulo et al.'s finding that lie-tellers took longer to respond when they had not prepared a response (see also Duñabeitia & Costa, 2015 and Spence et al., 2012 for more recent evidence of longer response latencies among lie-tellers than truth-tellers). In another recent meta-analysis exploring verbal cues to deception using computer software, lie-tellers' accounts were more simplistic than truth-tellers', as evidenced by the use of fewer words and diminished creativity (Hauch et al., 2015). Thus, lie-tellers have exhibited multiple cues indicative of increased cognitive complexity that distinguishes them from truth-tellers.

Lie-tellers have also been theorized to experience an increase in sympathetic nervous system arousal as a result of the emotions associated with lying (e.g., Zuckerman et al., 1981). Again, this would suggest that there should be identifiable differences in

arousal cues between lie- and truth-tellers. Brain imaging has shown that regions associated with emotional processing, such as the prefrontal cortex and amygdala, exhibited increased activity during deception (Abe et al., 2007). In observational studies, lie-tellers exhibited more dilated pupils, nervousness (tenseness, both overall and vocally), and fidgeting than truth-tellers (DePaulo et al., 2003). They also spoke in a significantly higher vocal pitch (DePaulo et al., 2003; Sporer & Schwandt, 2006). Based on these findings, it has been proposed that arousal cues can distinguish between lie- and truth-tellers. However, these arousal-based cues are limited in scope (i.e., there are fewer of them), particularly when compared to cognitive cues.

This brief overview of the literature on cues to deception might suggest that findings have been robust and unequivocal; however, that is not the case (see Sporer & Schwandt, 2006). For example, there was considerable variability in effect sizes across studies reported by DePaulo et al. (2003) and Hauch et al. (2015), as indicated by significant homogeneity statistics for many cues. In addition, in some studies, nervousness reportedly increased with deception (e.g., Strömwall et al., 2006), whereas in others, differences between lie- and truth-tellers were nonsignificant (e.g., Kassin & Fong, 1999). Recently, Luke (2019) argued that, while significant effects have been reported, they should be considered relatively small. That is, systematic methodological errors and omissions may have artificially inflated the already inconsistent effects.

Measuring Cognitive Load and Arousal

There is no single, established method to assess cognitive load and arousal during deception. This might not be surprising as the operationalization of these constructs is typically not uniform across studies. Instead, cognitive load and arousal have been

measured using a variety of methods, ranging from self-report to systematic behavioral observations. As such, a troublesome explanation for the variability in the diagnosticity of cues across studies reported by Luke (2019) is that it may be entirely a product of inconsistencies in assessment methods rather than the cues themselves, or a combination of both. For example, consider the cue ‘overall nervousness' from DePaulo et al.’s (2003) meta-analysis. Its homogeneity statistic was significant, indicating that the effect sizes across the 12 studies that included this cue were inconsistent. Of those 12 studies, seven were accessible via abstract only or entirely accessible through Google Scholar and PsycINFO inquiries for review here (i.e., Burgoon & Buller, 1994; Burgoon et al., 1996; DePaulo et al., 1983; DePaulo et al., 1982; Hocking & Leathers, 1980; Horvath, 1973; Kraut & Poe, 1980). Five employed observers alone to make behavioral ratings, one employed self-report and observers’ ratings, and one employed a behavioral measure of nervousness. However, the findings from the studies that were accessible showed that the effects were often in the same direction: lie-tellers self-reported greater nervousness, were perceived as more nervous, and had significantly different physiological responses compared to truth-tellers. Despite this, it is important to determine whether this discrepancy in measures is yielding inconsistencies in the diagnosticity of cues.

Self-report Measures

A commonly used measure of cognitive load and arousal in the deception detection literature is self-report. Other well-established subdisciplines within psychology have used inventories – such as the Minnesota Multiphasic Personality Inventory-2 (Butcher et al., 1989), Psychopathic Personality Inventory (Lilienfeld & Andrews, 1996), Perceived Stress Scale (Cohen et al., 1988), Rusbult Commitment Scale (Rusbult et al.,

1998), McGill Pain Questionnaire (Melzack, 1975), and Beck Depression Inventory (Beck et al., 1961) – extensively to successfully assess psychological constructs. However, no comparable, validated self-report measures have been broadly adopted by deception researchers.

Instead, deception researchers have made attempts to examine cognitive load and arousal using ad hoc self-report items. For example, in one study, participants were asked to lie or tell the truth and subsequently report their cognitive load (e.g., "Did lying require high mental effort?"; "Did you often think about the fact that you were having to lie?") and arousal (e.g., "Were you aroused when lying? [e.g., did your heart beat faster?]"; "Did you feel guilty when lying?") using three-item scales (Caso et al., 2005). In another, lie- and truth-tellers were asked to self-report their cognitive load (e.g., "How difficult was the interview?"; "To what extent did you [have] to concentrate during the interview?") and arousal ("To which extent did you feel upset during, or directly after the interview?"; "To which extent did you feel nervous during, or directly after the interview?") using five- and four-item scales, respectively (Ströfer et al., 2016). Researchers have even had lie- and truth-telling participants rate themselves on a 19-item scale assessing how cognitively taxed (e.g., "How hard did you have to pay attention to your behaviours?"; "How hard did you have to pay attention to the experimenter's behaviours?") and aroused (e.g., "How nervous were you when answering the interviewer's questions?"; "How guilty did you feel when you were answering the interviewer's questions?") they felt during a mock interview about a suspicious event (Elliott & Leach, 2016). As illustrated, self-report scales have not been consistent across studies.

Not only has this inconsistency hindered comparisons, but it has raised concerns about the parsimony and validity of measures. It is unclear which and how many items are required for accurate measurement. More importantly, it is unknown whether measures have even been assessing the same underlying constructs. For example, despite all purporting to study 'cognitive load,' some researchers have focused on executive function as a whole (e.g., Caso et al., 2005), whereas others have examined subcomponents, such as inhibition and planning (e.g., Elliott & Leach, 2016). Cues to deception that have emerged following the publication of DePaulo et al.'s (2003) meta-analysis have not all been adequately conceptualized, either. It has been hypothesized that lie-tellers think harder than truth-tellers as they formulate a response (Vrij et al., 2008). 'Thinking hard' has since been associated with cognitive load in studies that have employed observers to make appraisals of participants' behaviors (e.g., Evans et al., 2013; Mann & Vrij, 2006), but also those that have asked participants to rate their own cognitive effort (e.g., Elliott & Leach, 2016). However, it is unclear exactly what constitutes 'thinking hard.' Among many possibilities, it could encompass struggling to develop, remember, inhibit, and/or vocalize a response or even to understand the questions, instructions, and/or material that has been presented. Without a clear definition, there could be variability in terms of how participants interpret and respond to items. On the whole, the lack of standardized operational definitions for constructs and well-delineated items could undermine validity.

More generally, there are problems inherent in self-report. Participants may be unaware of and unable to report their internal processes, potentially skewing their accuracy when judging their own traits (see Nisbett & Wilson, 1977). If participants did

not know how they were feeling while lying, then the ability of researchers to adequately distinguish between lie- and truth-tellers based on self-reports would be severely limited. People may also attempt to manage impressions of themselves by responding or acting in a way they perceive as socially desirable (see Paulhus, 1984). For example, respondents could report feelings that they believed the researcher was expecting rather than what they experienced (e.g., stating that they felt aroused when they actually did not). Research has shown that there is a pancultural stereotype of deception (e.g., most people believe that lie-tellers avoid eye contact, are nervous, and provide incoherent responses, among other indicators; The Global Deception Research Team, 2006). Participants may rely on these beliefs when responding to items intended to assess their feelings during deception. Thus, it is unclear whether self-reported cognitive load and arousal cues are reliable and valid.

Trained Coders' Observations

Objective approaches to measuring cues to deception could be preferable to self-report. Indeed, cues to deception have frequently been assessed by having two or more observers independently make judgments about a participant's behaviors (e.g., Strömwall et al., 2006; Vrij et al., 2000). Observers have typically been trained to identify specific cues by researchers, and range in lie detection experience from university students (e.g., Mann et al., 2002) to police officers (e.g., Mann & Vrij, 2006). There has been no standard approach across studies regarding which, or how many, cues have been coded. For example, in one study, observers were asked to rate ten different cues, including both cognitive (e.g., pauses) and arousal (e.g., gaze aversion, illustrators) cues (Strömwall et al., 2006). In two other studies, they were asked to rate nine cognitive (e.g., number of

words spoken) and arousal (e.g., pitch, speech errors) cues (Sporer & Schwandt, 2006), and 13 cognitive (e.g., pauses) and arousal (e.g., blinking, head nods) cues (Mann & Vrij, 2006). Each of these cues has also been assessed using various approaches, such as having observers assess the frequency (e.g., number of pauses; Vrij et al., 2000), duration (e.g., length of pauses; Vrij et al., 2000), and/or degree of particular behaviors (e.g., plausibility; Leal et al., 2010). The variability in the type, number, and measurement of cues has complicated the establishment of reliability and validity of trained coders' assessments across studies.

Addressing this lack of standardization, Evans and colleagues (2013) developed the Psychologically Based Credibility Assessment Tool (PBCAT). The measure included 11 items and could be used by minimally-trained observers to evaluate a range of behaviors exhibited by lie- and truth-tellers. While watching interviews, observers used a paper scoring card to rate five items (i.e., auditory details, spatial details, temporal details, admitted lack of memory, and spontaneous corrections) on scales from 0 (not present) to 2 (frequent or present [3 or more times]), and six items (i.e., overall quantity of details/talking time, contradictions/plausibility, thought hard, nervousness, negativity/complaints, and rate of speech) on continuous nine-point scales ranging from -2 to 2. Among the 11 items on the scale, two were directly relevant to the cognitive load and arousal constructs assessed in the current study. Specifically, observers were asked how hard participants appeared to be thinking on a nine-point scale ranging from 'did not think hard' to 'thought extremely hard,' and how nervous participants appeared to be on a nine-point scale ranging from 'extremely relaxed/comfortable' to 'extremely tense/nervous.' Support for the inclusion of these items came from research showing that

lie-tellers think harder than truth-tellers (Vrij et al., 2008) and are believed to appear more nervous than truth-tellers (The Global Deception Research Team, 2006). Furthermore, both items had often been employed by observers in deception research (e.g., Landström et al., 2005; Mann & Vrij, 2006; Vrij et al., 2001). Importantly, the items differentiated between lies and truths in a direct test of the PBCAT (Evans et al., 2013). However, it remains unclear whether tools such as the PBCAT can be considered the 'gold standard' when assessing cognitive load and arousal because there are few points of comparison available to determine their validity more generally. That is, no research has compared the items to other measures of the same constructs.

Behavioral Measures

Other objective measures do exist: numerous studies have assessed behavioral indicators of cognitive load and arousal (Vrij, 2008). One measure of cognitive load has been response latency (i.e., the amount of time between an interviewer completing a question and the speaker beginning to respond; e.g., DePaulo et al., 2003; Sporer & Schwandt, 2006; Vrij et al., 2001). To assess response latency, two or more coders have typically watched videotaped interviews and indicated the time when each question ended and when the participant began speaking, thus capturing the silent pause in between (Vrij et al., 2000). All latencies have then been averaged. Although time-consuming, coding response latency has been used as a proxy for how long participants are taking to think about their responses and, thus, their cognitive load (Sporer & Schwandt, 2006).

Physiological measures of sympathetic nervous system arousal, such as the polygraph, have also long been used in deception detection research (Synnott et al.,

2015). A polygraph test consists of an examiner asking a suspect questions in an effort to elicit physiological arousal. Although the questions asked during a polygraph test could be structured in a variety of ways, the most common technique has been the comparison question technique (CQT), in which both relevant questions (i.e., questions specifically pertaining to the event under investigation) and comparison questions (i.e., questions not directly related to the event designed to stimulate arousal) are asked (Iacono & Ben-Shakhar, 2019; Saxe et al., 1985). During the test, physiological changes in response to both relevant and comparison questions have been measured using three components: electrodermal activity, heart rate, and respiration (Lykken, 1974; Saxe et al., 1985). The average of these three measures has typically been calculated to determine physiological arousal; however, it has been argued that each measure taps into different underlying processes (Vrij, 2008). Of the three, electrodermal activity – assessed by the perspiration obtained by connecting participants' fingers to electrodes – has been reported to be the most popular and reliable (Figner & Murphy, 2011; Lykken & Venables, 1971; Podlesny & Raskin, 1977). However, it is important to note that using physiological measures of arousal also requires sufficient resources, including the time to implement the instrument and train operators, as well as the necessary hardware and software required to collect data. Thus, it may not be the most cost- or time-efficient method when compared to self-report and trained coders' observations.[1]

[1] There have been concerns raised about the polygraph, but they have focused primarily on its diagnosticity of deception and vulnerability to countermeasures (see Iacono & Ben-Shakhar, 2019) rather than each component's ability to measure arousal. Because the current study was concerned solely with the measurement of physiological arousal, the overarching debate surrounding the reliability and validity of the polygraph was not relevant.

Comparing Measures

Few researchers have compared different ways of operationalizing and measuring cognitive load and arousal empirically. When approaches have been compared within individual studies, contradictions emerged. For example, lie-tellers reported having more difficulty responding to questions than truth-tellers, but veracity had no influence on coded cognitive cues to deception (Elliott & Leach, 2016). Similarly, lie-tellers reported feeling more nervous and strained than truth-tellers, yet veracity did not impact trained coders' observations of various arousal cues (Strömwall et al., 2006). These findings suggest that the operationalization of constructs will determine whether differences between groups are found. It is worth noting that, in their meta-analysis, DePaulo et al. (2003) compared objective measurements (e.g., counts/durations) to subjective measurements (e.g., trained coders' observations) of five cues, and found that effects were stronger for eye contact and facial pleasantness when measured subjectively than objectively. However, this analysis was limited by the available research at the time, as well as the operationalizations of the constructs of interest (e.g., thinking hard was not yet considered an important cognitive variable and, thus, was not included in the analysis). Regardless, an implicit assumption in the deception literature has been that all measures of cognitive load and arousal are valid (i.e., that all measures accurately and similarly measure cognitive load and arousal). To date, that has not been tested empirically.

Current Study

I tested the equivalence of commonly used measures of cognitive load and arousal. Data were collectively analyzed from three previous studies (Lahay et al. [*in preparation*], Woolridge et al. [2020], and Elliott & Leach [2016], henceforth referred to

as Studies A, B, and C, respectively). Across these studies, three types of measures of lie- and truth-tellers' cognitive load and arousal were employed: self-report, trained coders' observations, and quantification of objective behaviors. Specifically, participants were asked to report their cognitive load and arousal across several questionnaire items, participants' cognitive load and arousal (i.e., thinking hard and nervousness) were rated by trained coders, and behavioral measures – namely, response latency and sympathetic nervous system arousal[2] – were collected. These measures were selected because they have frequently been used in the deception detection literature to measure cognitive load and arousal (see Vrij, 2008).

This comprehensive approach allowed for the development of a multitrait-multimethod (MTMM) matrix and the establishment of the reliability and construct validity (including convergent and discriminant validity) of these measures (see Campbell & Fiske, 1959). To generate an MTMM matrix, Campbell and Fiske (1959) asserted that researchers needed to use multiple methods and assess multiple traits (i.e., constructs). The goal was to ensure that correlations were due to trait variance as opposed to method variance (i.e., variance that resulted from the method used rather than the constructs being evaluated; Podsakoff et al., 2003). Campbell and Fiske (1959) laid out four requirements to establish construct validity: (1) convergent validity values must be significant; (2) convergent validity values must be greater than the bordering correlations between different methods and traits; (3) convergent validity values must be greater than the values that share different traits but the same method, and these values should be uncorrelated; and, (4) there should be a recognizable pattern among the correlations

[2] An objective measure of sympathetic nervous system arousal was only employed in Study A.

within and between each of the heterotrait-hetero/monomethod triangles. Thus, here, each of the constructs (traits) and associated methods were analyzed using an MTMM analysis, paying particular attention to the four requirements outlined by Campbell and Fiske.

Hypotheses

I hypothesized that cognitive load could be equally measured using self-reports, trained coders' observations of thinking hard, and average response latency, and arousal could be equally measured using self-reports, trained coders' observations of nervousness, and average skin conductance level, as assessed by an MTMM analysis. Given that "reliability is a necessary but not sufficient condition for validity" (Picardi & Masick, 2013, p. 44), I hypothesized that the reliability values for each of these measures would be the largest values in the matrix when they were available, and that the measures would be reliable. Regarding convergent validity, I hypothesized that the interrelationship between each of the measures for each construct would be significant. In terms of discriminant validity, I hypothesized that the validity values would be greater than the correlations between dissimilar traits and methods and the correlations between dissimilar traits but the same method (i.e., measures intended to assess the same construct should be more strongly correlated than any other measures). Finally, I hypothesized that there would be a noticeable pattern to the relationship among traits in each of the heterotrait-hetero/monomethod triangles (e.g., if one correlation were twice as large as another within one triangle, then we would expect to see that same pattern in all other triangles). These hypotheses were largely guided by the requirements for establishing construct

validity outlined by Campbell and Fiske (1959), as well as the assumption in the deception detection literature that these measures were equivalent.

Method

Data were collected from three previous studies (i.e., Studies A, B, and C), all of which shared a similar task: the Suspicious Event Paradigm (Elliott & Leach, 2016). Participants were either instructed to lie or tell the truth about witnessing an event. That is, all were asked to describe an innocuous office scene, even though half had actually viewed evidence of a bomb plot. Participants' motivation was increased by telling them that they would receive bonus financial compensation for being convincing (Studies A, B, and C) or have to complete an undesirable task (i.e., write an essay) if they were not convincing (Study A).[3] All participants were interviewed about what they had seen using a structured interview that progressed from information-gathering questions (e.g., "What was on the wall?" "What was marked on the calendar?") to accusatorial questions (e.g., "Where was the gun?" "Are you lying to me?"). Interviews were videotaped in their entirety. In Study A, physiological data were collected during the entire interview. At the end of the interview, everyone completed self-report measures regarding their feelings during the interview and the interview more generally.

Design

The data from Studies A, B, and C were collectively analyzed as secondary data. Once all videotaped interviews were collected, observational ratings and calculations of objective response latency as a measure of cognitive load were made. All data were then collectively analyzed using a multitrait-multimethod analysis (Campbell & Fiske, 1959).

[3] In fact, all participants were told that they had been convincing.

Participants

As per a two-tailed *a priori* G*Power analysis (Faul et al., 2007) for a bivariate normal model using a medium effect size (0.25), the required sample size was 164 participants to achieve 90% power. Data from 238 participants were collected and analyzed. The number of participants exceeded the required sample size because secondary data were used, constraining the current study to the existing number of participants. Participants in Studies A (n = 83), B (n = 118), and C (n = 37) were undergraduate students from two mid-sized universities located in Canada, as well as community members surrounding these institutions.

Materials

Self-report Measures

In all studies, participants reported the cognitive load and arousal that they experienced during the interviews (see Appendices A and B).[4] Specifically, they responded to eight cognitive load items (i.e., how difficult it was to answer and understand the questions, how much they thought about and planned their answers, how much they paid attention to their own and the interviewers' behaviors, how hard it was to remember their answers, and how long they thought about their answers) and nine arousal items (i.e., how nervous, excited, guilty, surprised, ashamed, afraid, anxious, negative, and emotional they felt). All items were reported on scales from "not at all" to "extremely" (i.e., from 1 to 9 in Study A and from 1 to 10 in Studies B and C). The eight cognitive load items produced a Cronbach's alpha of .857, and the corrected item-total correlations were all acceptable. A preliminary analysis revealed that the nine arousal

[4] The materials for this study can also be accessed at: https://osf.io/n3e4x.

items produced a Cronbach's alpha of .825. However, upon examination of the corrected item-total correlations, the 'excited' item produced a value of .022. In addition, the Cronbach's alpha if item deleted value for the 'excited' item revealed that the alpha would increase to .858 if that item were to be removed. Thus, the item was removed from all subsequent analyses, bringing the eight-item arousal scale's Cronbach's alpha to .858. The mean score for each construct was then calculated to create the "self-reported cognitive load" and "self-reported arousal" variables.

Additionally, participants responded to one item pertaining to motivation. They were asked how motivated they felt to convince the interviewer(s) they were telling the truth, on a scale from 1 (not at all) to 9 (extremely; Study A) or from 1 (not at all) to 10 (extremely; Studies B and C). This item was included in the MTMM analysis to assist in establishing discriminant validity, as motivation should not be related to either cognitive load or arousal.

Trained Coders' Observations

Four independent coders were trained to identify the variables of interest (see Appendices C and D for coding materials used by the coders). Recorded interviews were randomly assigned to coders and counterbalanced. Each item – "thinking hard," "nervousness," "excitement," and "motivation" – was coded by two blind coders. Specifically, they indicated how hard each participant appeared to be thinking (from 1 [Did not think hard] to 9 [Thought extremely hard]), how nervous each participant appeared to be (from 1 [Extremely relaxed/comfortable] to 9 [Extremely tense/nervous]), how excited they appeared to be (from 1 [Not at all excited] to 9 [Extremely excited]), and how motivated the participant appeared to be to convince the interviewers they were

being honest (from 1 [Not at all motivated] to 9 [Extremely motivated]). The "thinking hard" and "nervousness" items were adapted from the PBCAT (Evans et al., 2013), and the operationalizations for each variable that were provided in the PBCAT coding materials were used.

Inter-rater reliability was assessed on 100% of the coded items for the combined studies using one-way intraclass correlation coefficients (ICCs). Discrepancies were resolved through discussion. Reliability was moderate for trained coders' observations of thinking hard, nervousness, and motivation (for cutoff values, see Koo & Li, 2016). Specifically, the average measure one-way ICCs were .600 for thinking hard (F (237, 238) = 2.497, $p < .001$, 95% CI [.483, .690]), .684 for nervousness (F (237, 238) = 3.164, $p < .001$, 95% CI [.592, .755]), and .634 for motivation (F (237, 238) = 2.736, $p < .001$, 95% CI [.528, .717]). Due to the removal of excitement from the self-report scale of arousal, coders' ratings of excitement were not analyzed further (i.e., the item was omitted from all subsequent analyses).

Behavioral Measures

Four independent, trained coders used behavioral coding software (i.e., Datavyu; see Datavyu Team, 2014) to assess response latency (i.e., an objective measure of cognitive load; see Appendix E for the response latency coding manual). Each recorded interview was coded by two people who were blind to condition and my hypotheses. For each question, they coded the amount of time between when the interviewer stopped speaking and when a participant started speaking. The average latency across questions was then calculated to produce an "average response latency" in milliseconds (ms) for each participant. Inter-rater reliability was calculated on 100% of the interviews for the

combined studies using one-way ICCs, and discrepancies were resolved by a third coder.[5] Reliability was excellent for response latency, with an average measure one-way ICC of .975, F (5503, 5504) = 39.878, $p < .001$, 95% CI (.974, .976).

In Study A, participants were connected to a device designed to measure their electrodermal activity (EDA; i.e., the Biopac MP160 system). EDA is a combination of skin conductance level (SCL) and skin conductance responses (SCRs; Braithwaite et al., 2015). Using exosomatic DC recording – a small electric current applied to the skin (Boucsein et al., 2012) – the device recorded how active participants' eccrine glands were in response to physiological arousal (Biopac Systems, 2015). The participant's EDA information was wirelessly transmitted through the Biopac system to the software, AcqKnowledge 5, which recorded any changes in their EDA at 2000Hz (2000 samples/second) as a continuous line (an indication of microsiemens [μS], which is a measure of electrical conductance whereby the normal range of electrical conductance is 2-20 μS; Biopac Systems, 2015). Skin conductance data were collected from 83 participants, which were then processed for analysis.

I assessed participants' average SCL – an overall indication of sympathetic nervous system arousal (Boucsein et al., 2012). All files were pre-processed using AcqKnowledge to down-sample each file to a rate that was less computationally intensive. Based on the existing literature (e.g., the Nyquist theorem; see Braithwaite et al., 2015), a sampling rate of 200Hz (200 samples/second) was selected. After down-sampling, the SCRs were populated using AcqKnowledge. The average of several intervals across each waveform was taken (Figner & Murphy, 2011); specifically, every

[5] This strategy was used because discrepancies could not be resolved through discussion due to the unavailability of one of the coders.

2-second interval before the onset of each SCR was averaged. An interval of 2 seconds was chosen because interview questions were occasionally presented in a staccato rhythm (with 2 seconds being a common latency), and the goal was to avoid overlapping SCRs. In addition, SCL is commonly considered to be the waveform without SCRs, and so excluding those from the calculation of the average was necessary (Braithwaite et al., 2015).

Results

Descriptive statistics for each study are presented in Table 1.[6] For the multitrait-multimethod analysis, aggregate results are presented. Analyses were conducted on each study separately, as well. However, the pattern of results was similar across studies, so the matrices from the individual studies have been omitted for the sake of brevity.[7]

[6] Descriptive statistics are presented by study as the scale used varied between them.
[7] Analyses for each study are available upon request.

Table 1

Mean Cognitive Load, Arousal, and Motivation Scores

	Study A			Study B			Study C		
	M	***SD***	***N***	***M***	***SD***	***N***	***M***	***SD***	***N***
Self-report scales									
Cognitive load	3.76	1.41	83	4.60	1.84	118	4.11	1.80	37
Arousal	3.63	1.59	83	4.41	1.95	118	4.17	1.88	37
Motivation	6.36	2.14	83	6.45	2.49	118	6.54	2.85	37
Trained coders' observations									
Thinking hard	2.49	1.16	83	3.03	1.23	118	2.27	1.48	37
Nervousness	3.05	1.45	83	2.31	1.09	118	3.78	1.70	37
Motivation	3.42	1.32	83	3.72	1.47	118	4.00	2.01	37
Behavioral measures									
Average RL	1018.73	448.74	83	1699.19	967.02	118	1660.56	746.83	37
Average SCL	10.48	4.57	83	-	-	-	-	-	-

Note. The self-report measures and trained coders' observations were rated on a scale from 1-9 (with higher scores indicating a greater amount of the cue) in Study A. The self-report measures were rated on a scale from 1-10 and the trained coders' observations were made on a scale from 1-9 (with higher scores indicating a greater amount of the cue) in Studies B and C. RL represents response latency and SCL represents skin conductance level.

Multitrait-Multimethod Analysis

Bivariate correlations were performed to assess the convergent and discriminant validity of the measures (see Table 2). Campbell and Fiske's (1959) MTMM analysis was used to evaluate the matrix; that is, I compared the results to each of the four criteria that they delineated. The majority of the reliability values, denoted by the values on the diagonal, were the largest in the matrix. Specifically, the Cronbach's alpha values for the two self-report measures were high, thus providing evidence for internal consistency of

those scales. The ICCs for trained coders' observations and response latency were primarily acceptable based on cutoff values (see Koo & Li, 2016), although the ICCs for trained coders' observations of thinking hard (r = .600) and motivation (r = .634) were less than the correlation between self-reported arousal and self-reported cognitive load (r = .665).

Table 2

Multitrait-Multimethod Matrix

Method / Trait		**A_1**	**B_1**	**C_1**	**A_2**	**B_2**	**C_2**	**A_3**	**B_3**
Self-report scales									
Cognitive load	**A_1**	(.857)							
Arousal	**B_1**	.665**	(.858)						
Motivation	**C_1**	.044	.079	( )					
Trained coders' observations									
Thinking hard	**A_2**	*.219***	-.002	.021	(.600)				
Nervousness	**B_2**	.060	*-.046*	-.020	.254**	(.684)			
Motivation	**C_2**	-.211**	-.083	*.094*	-.081	.016	(.634)		
Behavioral measures									
Average RL	**A_3**	*.178***	.050	.015	*.560***	.155*	-.055	(.975)	
Average SCL	**B_3**	.007	*-.041*	-.113	-.146	*-.045*	.203	-.079	( )

Note. Cronbach's alpha values and intraclass correlation coefficients are reported in parentheses. The remaining values are Pearson correlation coefficients representing the intercorrelations between constructs and methods. Convergent validity values are italicized. RL represents response latency and SCL represents skin conductance level. Letters represent the constructs (i.e., A = cognitive load; B = arousal; C = motivation), whereas associated subscripts represent the measure being used (i.e., 1 = self-report; 2 = trained coders' observations; 3 = behavioral measures).

* $p < .05$. ** $p < .01$.

Hypothesis #1: Convergent Validity Values Must be Significant

All cognitive load measures were positively related. Specifically, the correlation between trained coders' observations of thinking hard and average response latency (r = .560, p < .001), between self-reported cognitive load and trained coders' observations of thinking hard (r = .219, p = .001), and between self-reported cognitive load and average response latency (r = .178, p = .006) were all significant. Arousal measures all showed nonsignificant relationships with each other. Specifically, the correlation between self-reported arousal and trained coders' observations of nervousness (r = -.046, p = .484), between self-reported arousal and average skin conductance level (r = -.041, p = .711), and between trained coders' observations of nervousness and average skin conductance level (r = -.045, p = .684), were all nonsignificant. Finally, the correlation between self-reported motivation and trained coders' observations of motivation was nonsignificant (r = .094, p = .150). Therefore, Hypothesis #1 was only supported for the measures of cognitive load.

Hypothesis #2: Convergent Validity Values Must be Greater than Their Adjacent Heterotrait-Heteromethod Values

Hypothesis #2 was supported for cognitive load measures: the correlations between self-reported cognitive load and trained coders' observations of thinking hard (r = .219), self-reported cognitive load and average response latency (r = .178), and trained coders' observations of thinking hard and average response latency (r = .560) were all greater than their neighboring heterotrait-heteromethod values. Hypothesis #2 was not supported for motivational or arousal measures, however. Specifically, the correlation between self-reported motivation and trained coders' observations of

motivation (r = .094) was less than one of its adjacent heterotrait-heteromethod values. Similarly, I failed to find support for the hypothesis when examining the correlations between self-reported arousal and trained coders' observations of nervousness (r = -.046), self-reported arousal and average skin conductance level (r = -.041), or trained coders' observations of nervousness and average skin conductance level (r = -.045). In sum, Hypothesis #2 was only supported for measures of cognitive load.

Hypothesis #3: Convergent Validity Values Must be Greater than the Heterotrait-Monomethod Values

Although two of the convergent validity values were among the largest when compared to the heterotrait-monomethod correlations, no convergent validity values were greater than *all* of the heterotrait-monomethod correlations in the matrix. The correlation between self-reported cognitive load and self-reported arousal (r = .665) was larger than all other correlations. The correlation between trained coders' observations of thinking hard and average response latency (r = .560) was second only to the correlation between self-reported cognitive load and self-reported arousal.

Also, theoretically unrelated traits measured by the same method should not have been as highly correlated as traits expected to be related measured by the same method (Campbell & Fiske, 1959). Of course, this assumes that cognitive load, arousal, and motivation are truly independent constructs, as posited in deception theories (e.g., DePaulo et al., 2003). All but two of the heterotrait-monomethod correlations were nonsignificant. In particular, the correlations between self-reported cognitive load and self-reported arousal (r = .665), and between trained coders' observations of thinking hard and trained coders' observations of nervousness (r = .254), were significant

(both *p*s < .001), suggesting that shared method variance may have been contributing to the self-report scores and trained coders' observations. Therefore, Hypothesis #3 was not supported for any of the convergent validity values.

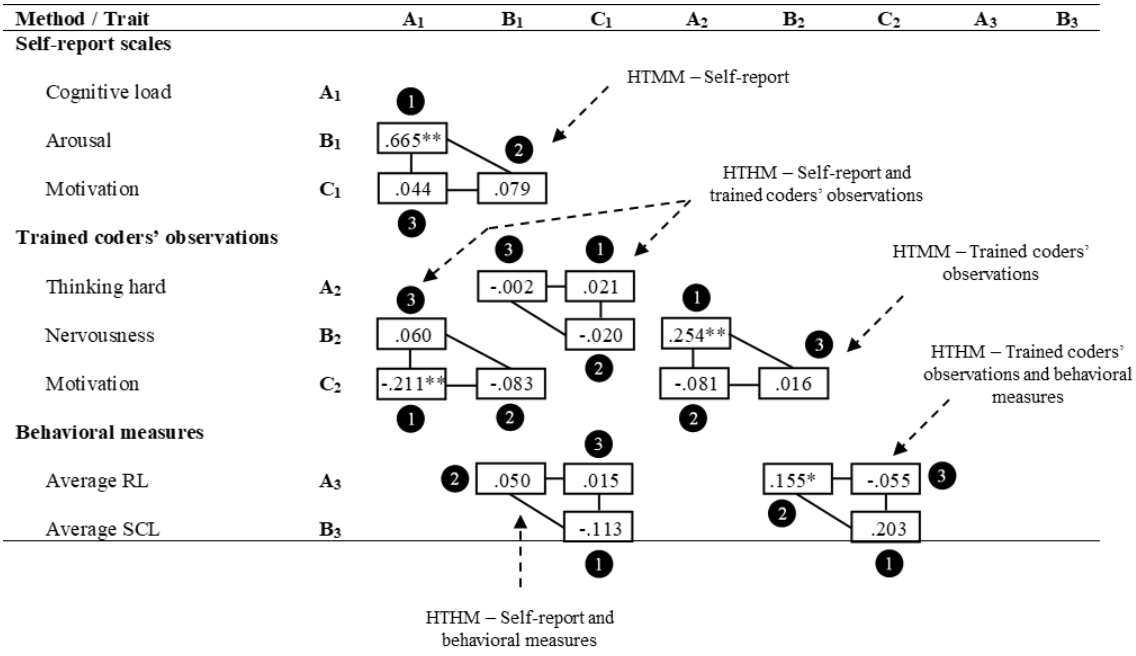
Hypothesis #4: There Must be an Observable Pattern in Each of the Heterotrait-Heteromethod and Heterotrait-Monomethod Triangles

Within complete triangles (i.e., those that contained three values and therefore three points of comparison for other triangles), there was not a distinguishable pattern (see Figure 1 for a visual depiction). In the heterotrait-monomethod triangle for self-report, the correlation between self-reported cognitive load and arousal (r = .665) was the largest, followed by the correlation between arousal and motivation (r = .079) and then cognitive load and motivation (r = .044). In contrast, while the pattern in the heterotrait-heteromethod triangles for self-report and trained coders' observations was identical, this pattern differed from that in the heterotrait-monomethod triangle for self-report. Specifically, the correlation between self-reported cognitive load and trained coders' observations of motivation (r = -.211) was larger than the correlations between self-reported arousal and trained coders' observations of motivation (r = -.083) and between self-reported cognitive load and trained coders' observations of nervousness (r = .060). For Study A, physiological data was also available, thus creating a full heterotrait-heteromethod triangle for self-report and behavioral measures and for trained coders' observations and behavioral measures. The pattern was similar in the heterotrait-heteromethod triangle for self-report and behavioral measures and the heterotrait-heteromethod triangle for trained coders' observations and behavioral measures; however, the pattern differed in the heterotrait-monomethod triangle for self-report, the

heterotrait-heteromethod triangles for self-report and trained coders' observations, and the heterotrait-monomethod triangle for trained coders' observations. Therefore, while there were *some* commonalities within the matrix, there were four distinct patterns. As such, there was no clear pattern of trait interrelationship between all of the heterotrait-heteromethod and heterotrait-monomethod triangles, indicating that Hypothesis #4 was not supported.

Figure 1

Complete Triangles Isolated for Pattern Identification



Note. HTMM represents a heterotrait-monomethod triangle and HTHM represents a heterotrait-heteromethod triangle. Numbers 1 through 3 represent the descending magnitude of each correlation within each triangle, with 1 being the largest and 3 being the smallest. Solid lines depict each triangle, while dashed lines lead to a description of each triangle.

Discussion

I used a multitrait-multimethod (MTMM) analysis (Campbell & Fiske, 1959) to assess the construct validity of common measures of cognitive load and arousal in deception research. I combined data from three studies, each of which utilized at least two of three measures of cognitive load and arousal, into a single analysis pipeline. Using the criteria outlined by Campbell and Fiske for analyzing the MTMM matrix produced mixed results regarding the reliability and construct validity of the measures. Specifically, all measures were reliable. Measures of cognitive load showed some evidence of construct validity, although measures of arousal did not. The findings from the current study may help to explain the lack of robustness and consistency of cue findings (Luke, 2019), which has been suggested to be due to the inconsistency in measurements used across studies.

I proposed several hypotheses pertaining to the reliability and construct validity of the measures. I predicted that all of the measures would be reliable and the reliability values would be the largest in the matrix. The analysis of the MTMM matrix revealed that the self-report measures of cognitive load and arousal were reliable. Likewise, trained coders' observations of thinking hard, nervousness, motivation, and response latency were all reliable, as evidenced by acceptable reliability scores for those measures (see Koo & Li, 2016). However, not all of the reliability values were the largest in the matrix; specifically, the values for trained coders' observations of thinking hard and motivation were slightly less than the correlation between self-reported arousal and cognitive load. This is somewhat problematic given that measures need to be reliable first and foremost (Cronbach, 1951); that is, to be valid, a measure needs to be reliable. Some

of the reliability values being less than other values in the matrix may indicate that those specific measures are flawed. Despite this, the findings suggest that most of the measures were consistent, which is unique in that I was able to assess reliability across three independent studies.

The findings aligned with other studies that have shown acceptable to high reliability for self-report (e.g., Caso et al., 2005), trained coders' observations (e.g., Vrij et al., 2000), and behavioral measures (e.g., response latency; Vrij et al., 2000). However, there is a noticeable range in the reliability values reported. For example, whereas the Cronbach's alpha that I reported for the cognitive load scale was similar to that in Caso et al. (2005), the Cronbach's alpha for arousal was much more robust in the current study. Vrij et al. (2000) reported greater than 90% agreement for all of the cues that observers coded, yet trained coders' observations in the current study were in the 60% to 70% agreement range (excluding response latency). In other studies, reliability was not reported (e.g., Elliott & Leach, 2016; Evans et al., 2013), making it difficult to verify whether their measures of cognitive and arousal cues to deception were reliable.

Based on Campbell and Fiske's (1959) four criteria, I hypothesized that the validity values would be significant; this was partially supported. Specifically, convergent validity was established for the measures of cognitive load, as evidenced by significant correlations between each measure. Similar to other studies, the self-report measure in the current study contained several items intended to assess cognitive load (e.g., Caso et al., 2005; Elliott & Leach, 2016), and the trained coders used the same 'thinking hard' item included on the PBCAT (Evans et al., 2013). Given previous criticisms of subjective measures (e.g., Nisbett & Wilson, 1977), the correlation between

all measures might be surprising. Yet, in a deception study in which the number of details provided by an interviewee was both objectively and subjectively evaluated, both types of ratings were highly correlated (Ewens et al., 2016). Likewise, my finding suggests that common measures of cognitive load in deception research may be assessing a similar construct.

Contrary to my hypothesis, convergent validity was not established for any measures of arousal. This finding suggests that the measures of arousal employed in this study may have been assessing different underlying processes. There have been criticisms of measurements of arousal in the literature (e.g., Synnott et al., 2015; Vrij, 2008); for example, participants might have altered their responses to be more socially desirable (Paulhus, 1984), and lie-tellers may have expected to feel aroused based on pancultural stereotypes associated with lying (e.g., The Global Deception Research Team, 2006). This could account for why their reports did not map onto physiological responses assessed with more objective measures (Pennebaker, 1999). Moreover, the variability in arousal cue findings identified by DePaulo et al. (2003) might suggest that arousal may be too broadly defined to measure reliably. Even each component of the polygraph has been suggested to measure a different dimension of physiological arousal, such that some measures may reflect higher or lower arousal overall compared to other measures (Vrij, 2008). Thus, it stands to reason that arousal would not have been fully encapsulated by a single cue (e.g., nervousness) or component (e.g., EDA) in my (and others') research.

I had also predicted that there would be evidence of discriminant validity. In particular, I hypothesized that the validity values would be greater than those between dissimilar methods and/or traits, that measures of cognitive load and arousal would be

uncorrelated with similar measures of different traits, and that there would be an identifiable pattern of trait interrelationship within the heterotrait-hetero/monomethod triangles. Only the correlations between trained coders' observations of thinking hard and average response latency, self-reported cognitive load and trained coders' observations of thinking hard, and self-reported cognitive load and average response latency were greater than their neighboring correlations. There were no validity coefficients that were greater than *all* of the heterotrait-monomethod correlations. Additionally, there were significant correlations between different traits measured by the same method and no pattern of correlations in the matrix. Many of the correlations between my traits of interest and motivation were nonsignificant, thus providing some evidence for discriminant validity. However, there was a significant correlation between self-reported cognitive load and trained coders' observations of motivation. This finding is concerning given that cognitive load was not expected to vary with motivation, as they were believed to be independent constructs. As such, the majority of criteria required by Campbell and Fiske (1959) for discriminant validity were not met for both cognitive load and arousal, and traits that should be unrelated varied together to some extent.

These findings suggest that the favorable convergent validity correlations among measures of cognitive load were possibly influenced by shared method variance. In other words, it may be that the unique features of each method were artificially inflating the convergent validity values by systematically contributing to the variance among them (Maul, 2013; Podsakoff et al., 2003). In addition, cognitive load and arousal might not be independent. Indeed, some studies have shown that cognitive load and arousal vary together (e.g., Leal et al., 2008). It could be that both are related to an underlying

construct. For instance, it is possible that emotion was the overarching construct, as theories have suggested that it is comprised of physiological arousal and cognition (e.g., Schachter & Singer, 1962). There are also some limitations to the current study that may help to explain this pattern of results.

Limitations and Future Directions

Campbell and Fiske (1959) noted that a reduction in the sample size across traits could be problematic for the interpretation of the MTMM matrix. Physiological data were only collected in one of the studies (Study A), thereby limiting my ability to compare measures for all participants or between studies. Ideally, I would have had all participants complete every single measure of cognitive load and arousal. However, I was limited by the use of secondary data. Although a relatively small sample completed the measure of physiological arousal, the other cell sizes exceeded the threshold required according to my *a priori* power analysis. The findings related to behavioral measures of arousal in this study, therefore, might be underpowered and could merit additional research, but the remainder of the findings were robust.

In addition, the same deception paradigm was used in all of the studies examined here. It may not have evoked as much cognitive load or arousal as other 'high-stakes' paradigms (e.g., Porter & ten Brinke, 2010) or produced sufficient variance. Researchers might consider evaluating whether findings from the current study generalize to other deception paradigms. Yet, in a meta-analysis, behavioral differences between lie-tellers and truth-tellers were similar under both low and high motivation conditions (Hartwig & Bond, 2014). Regardless of motivation, cognitive load and arousal cues should continue

to be more salient among lie-tellers than truth-tellers, and different measures should similarly identify those differences if they show construct validity.

Finally, there may be limitations to the MTMM matrix. Criticisms of the traditional MTMM analysis by Campbell and Fiske (1959) have included that it faces measurement error, lacks the support of statistical analyses, and cannot decompose the results into the unique influence of method and trait variance (Koch et al., 2020; Schmitt & Stults, 1986). In addition, the traditional MTMM analysis has not been empirically supported and the criteria are somewhat arbitrary. Instead, it has been suggested that researchers use methods, such as confirmatory factor analysis (CFA), to analyze the matrix (e.g., Eid, 2000; Kenny & Kashy, 1992; Marsh, 1989); however, those who have suggested this approach have also identified potential problems with certain CFA techniques. Importantly, CFA – and structural equation modeling more generally – has been suggested to require a large sample size (Kline, 2016; Marsh, 1989), which is something deception researchers have struggled to attain (Luke, 2019). With that said, Campbell and Fiske's (1959) criteria alone have been considered diagnostic of construct validity (Lance et al., 2002; Marsh, 1989; Strauss & Smith, 2009), and their work stands as a pioneering publication on convergent and discriminant validity in psychological research (e.g., Marsh, 1989). Although researchers might consider taking the next step to CFA to look at the unique influence of method and trait variance, that was beyond the scope of this paper. Moreover, as Campbell and Fiske (1959) noted:

> Psychologists today should be concerned not with evaluating tests as if the tests were fixed and definitive, but rather with developing better tests. We believe that a careful examination of a multitrait-multimethod matrix will indicate to the

> experimenter what [their] next steps should be: it will indicate which methods should be discarded or replaced, which concepts need sharper delineation, and which concepts are poorly measured because of excessive or confounding method variance (p. 103).

I adopted a similar stance.

Implications

This study is one of the first empirical examinations of the equivalence of common measures of cognitive load and arousal within the deception detection literature. Although researchers have typically assumed that measurements of cognitive load and arousal are relatively interchangeable, the findings from the current study suggest that they are not. There was some evidence that cognitive load was comparably measured using self-reports, trained coders' observations, and behavioral measures. Thus, there is support for researchers comparing findings obtained using different methods and employing the most resource-efficient method when designing their own studies. However, there remain concerns associated with using self-reports (Nisbett & Wilson, 1977; Paulhus, 1984) and the evidence for the discriminant validity of measures in the current study was weak. Instead, researchers might rely on validated scales that observers can use, such as the PBCAT (Evans et al., 2013), instead of a more laborious technique, such as coding response latency. When removing the self-report measures from the matrix to more closely examine the relationship between trained coders' observations of thinking hard and average response latency, Hypotheses #1 through #3 were met for both; Hypothesis #4 was not. Based on this finding and the lack of research on the relationship

among these more objective assessments, additional research is needed before a particular approach is more widely endorsed to measure cognitive load.

In contrast, albeit reliable, common measures of arousal did not appear to be equivalent. Researchers should be using extreme caution when employing and interpreting arousal measures. Although it may be tempting for researchers to rely on more objective measures, such as the polygraph, these techniques have been met with criticism. In particular, researchers have warned against interpreting the results of a polygraph as an indication of guilt due to a lack of evidence demonstrating its validity (Iacono & Ben-Shakhar, 2019; Lykken, 1974). For the same reason, although Canadian police services are permitted to utilize the polygraph, evidence from this test is inadmissible in Canadian courts (*R v. Béland*, 1987; *R v. Oickle*, 2000). With this in mind, there is no definitive recommendation of which measure of arousal is empirically best. Specifically, it is unclear what the measures of arousal used in the current study were assessing given that they were seemingly unrelated to each other, which may serve as an explanation for the variability in arousal cues that has been reported previously (see DePaulo et al., 2003; Luke, 2019). My work suggests that results from different operationalizations of arousal should not be compared within or between studies (including within meta-analyses) without measures being validated.

The goal is not to highlight that my measures provide the most accurate assessment of cognitive load and arousal; there may be better measures available. Instead, I have taken a first step in assessing measures' construct validity and relationship to one another. This study underscores the importance of this approach: it may have been logical to presume that any two of the measures employed would be assessing the same

construct, yet the findings suggest that they were not. Indeed, the measures may have been assessing different components of cognitive load and arousal. For example, in the case of cognitive load, executive functioning can be further divided into processes, such as inhibition, planning, and working memory (Christ et al., 2009; Gombos, 2006). As such, the cognitive load measures used in this study may still be valuable, although simply could have been assessing one (or several) of these components. The self-report measures partially accounted for these various processes because they comprised several related items and the mean scores were analyzed. A comparison with a global measure of cognitive load and arousal would assist in determining whether individual items are mapping onto the constructs more broadly. In addition, 'thinking hard' has regularly been used and lauded in the literature as a measure of cognitive load (e.g., Elliott & Leach, 2016; Evans et al., 2013), yet this item has not been clearly defined and lacks empirical support. It may be useful to turn to the cognitive literature to develop a more robust understanding of how cognitive load presents or may be expected to present. Regardless, researchers cannot simply assume that all measures of cognitive load and arousal are the same, nor that they are using the best possible measure simply because it has been employed historically. They must ensure that the measures being used assess the construct of interest (Strauss & Smith, 2009). The deception literature has largely failed in this regard.

There are several modifications that researchers can make to their current approaches moving forward. My findings suggest that when a specific self-report measure of cognitive load and arousal is employed across three independent studies, similar results can be obtained. To reduce error and the inability to compare measures

across studies, deception researchers should consider validating and refining existing self-report measures of cognitive load and arousal rather than constantly developing new ones. Furthermore, it is not advisable to employ one measure of cognitive load and arousal or employ multiple measures but then discount the measure that did not 'work' (Campbell & Fiske, 1959). Rather, researchers should consider employing multiple measures to assess cognitive and arousal cues and then correlating those measures to confirm that they are measuring the construct of interest, as was done in the current study using an MTMM analysis (Campbell & Fiske, 1959). Researchers might also consider creating a latent factor for each of these constructs. If a cue is not performing as expected across studies, then it is important to examine methodological explanations. For example, the homogeneity statistic for the cue 'overall nervousness' reported in DePaulo et al.'s (2003) meta-analysis was significant. Upon closer examination, three different measures were employed in the included studies; my work suggests that the variability in results could have been due to these methods. To avoid this issue going forward, researchers should correlate measures and traits to determine whether variance is due to the cue itself or to the type of measurement used to assess that cue (Campbell & Fiske, 1959). In other words, the tools used to assess cues to deception, or any psychological construct for that matter, should be validated (Strauss & Smith, 2009).

As constructs, cognitive load and arousal have been common to many theoretical models of deception (DePaulo et al., 2003; Zuckerman et al., 1981). Both have been predicted to increase when a person is lying, thus giving rise to identifiable cues to deception. It may be problematic that researchers are continually refining these theories and developing training programs based on their findings when studies of cue

diagnosticity have been grounded in untested and largely unvalidated measures. This is not to say that researchers should stop assessing the diagnostic value of cues to deception or that existing cues are not diagnostic. Instead, it is to highlight that traditional cognitive and arousal-based measures may be tapping into different – and perhaps unrealized – constructs. There appear to be significant discrepancies in effect sizes for cues to deception across studies (see Luke, 2019), and, in the current study, empirical measurements of constructs yielded mixed results. In particular, my findings suggest that arousal, as theorized, poses challenges in terms of measurement and diagnosticity. In fact, some studies since the publication of DePaulo et al.'s (2003) meta-analysis have reported that lie-tellers do not actually display more nervousness than truth-tellers (e.g., Vrij & Fisher, 2020). Because models of deception should focus on cues that have been shown to be consistently reliable (DePaulo et al., 2003; Luke, 2019), the reconceptualization of arousal and its relationship to deception may be in order. Importantly, models of deception have begun to transition their focus to cognition due to concerns about arousal-based approaches (Vrij et al., 2017; Vrij et al., 2006; Vrij et al., 2010). My findings support this move.

Conclusion

I assessed the construct validity of common measures of cognitive load and arousal using a multitrait-multimethod analysis. Although equivalence has historically been assumed, it had rarely been tested empirically. I found some evidence for the construct validity of my measures. Specifically, the measures of cognitive load (i.e., self-reported cognitive load, trained coders' observations of thinking hard, and average response latency) exhibited convergent validity; however, there was no evidence for

discriminant validity. I found no evidence for the construct validity of the measures of arousal using the criteria outlined by Campbell and Fiske (1959). Generally, these results highlight that, although common measures of cognitive load and arousal are assumed to be equivalent, that is unlikely to be the case. Moreover, they offer preliminary support for the notion that discrepancies in the diagnosticity of cues to deception across studies (DePaulo et al., 2003; Luke, 2019) are attributable to the measures used rather than variability in the cues themselves. My work suggests that researchers should exercise caution when reaching conclusions about the roles of cognitive load and arousal in deception based on single measures because they may not be assessing intended constructs.

References

Abe, N., Suzuki, M., Mori, E., Itoh, M., & Fujii, T. (2007). Deceiving others: Distinct neural responses of the prefrontal cortex and amygdala in simple fabrication and deception with social interactions. *Journal of Cognitive Neuroscience, 19*(2), 287–295. https://doi.org/10.1162/jocn.2007.19.2.287

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004

Biopac Systems. (2015). *EDA introductory guide*. https://www.biopac.com/wpcontent/uploads/EDA-Guide.pdf

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., Dawson, M. E., Filion, D. L., & Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures (2012). Publication recommendations for electrodermal measurements. *Psychophysiology, 49*(8), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2015). *A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments*. University of Birmingham. https://www.birmingham.ac.uk/Documents/college-les/psych/saal/guide-electrodermal-activity.pdf

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory, 6*(3), 203–242. https://doi.org/10.1111/j.1468-2885.1996.tb00127.x

Burgoon, J. K., & Buller, D. B. (1994). Interpersonal deception: III. Effects of deceit on perceived communication and nonverbal behavior dynamics. *Journal of Nonverbal Behavior, 18*(2), 155–184. https://doi.org/10.1007/BF02170076

Burgoon, J. K., Buller, D. B., Floyd, K., & Grandpre, J. (1996). Deceptive realities: Sender, receiver, and observer perspectives in deceptive conversations. *Communication Research, 23*(6), 724–748. https://doi.org/10.1177/009365096023006005

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *The Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. University of Minnesota Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Caso, L., Gnisci, A., Vrij, A., & Mann, S. (2005). Processes underlying deception: An empirical analysis of truth and lies when manipulating the stakes. *Journal of Investigative Psychology and Offender Profiling, 2*(3), 195–202. https://doi.org/10.1002/jip.32

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex, 19*(7), 1557–1566. https://doi.org/10.1093/cercor/bhn189

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*(4), 385. https://doi.org/10.2307/2136404

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Datavyu Team (2014). *Datavyu: A video coding tool*. Databrary Project, New York University. http://datavyu.org.

DePaulo, B. M., Lanier, K., & Davis, T. (1983). Detecting the deceit of the motivated liar. *Journal of Personality and Social Psychology, 45*(5), 1096–1103. https://doi.org/10.1037/0022-3514.45.5.1096

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

DePaulo, B. M., Rosenthal, R., Green, C. R., & Rosenkrantz, J. (1982). Diagnosing deceptive and mixed messages from verbal and nonverbal cues. *Journal of Experimental Social Psychology, 18*(5), 433–446. https://doi.org/10.1016/0022-1031(82)90064-6

Duñabeitia, J. A., & Costa, A. (2015). Lying in a native and foreign language. *Psychonomic Bulletin & Review, 22*(4), 1124–1129. https://doi.org/10.3758/s13423-014-0781-4

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*(2), 241–261. https://doi.org/10.1007/BF02294377

Elliott, E., & Leach, A.-M. (2016). You must be lying because I don't understand you: Language proficiency and lie detection. *Journal of Experimental Psychology: Applied, 22*(4), 488–499. https://doi.org/10.1037/xap0000102

Evans, J. R., Michael, S. W., Meissner, C. A., & Brandon, S. E. (2013). Validating a new assessment method for deception detection: Introducing a Psychologically Based Credibility Assessment Tool. Jou*rnal of Applied Research in Memory and Cognition, 2*(1), 33–41. https://doi.org/10.1016/j.jarmac.2013.02.002

Ewens, S., Vrij, A., Leal, S., Mann, S., Jo, E., & Fisher, R. P. (2016). The effect of interpreters on eliciting information, cues to deceit and rapport. *Legal and Criminological Psychology, 21*(2), 286–304. https://doi.org/10.1111/lcrp.12067

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Figner, B., & Murphy, R. O. (2011). Using skin conductance in judgment and decision making research. In M. Schulte-Mecklenbeck, A. Kühberger, & R. Ranyard (Eds.), *A handbook of process tracing methods for decision research* (pp. 163–184). Psychology Press.

Gombos, V. A. (2006). The cognition of deception: The role of executive processes in producing lies. *Genetic, Social, and General Psychology Monographs, 132*(3), 197–214. https://doi.org/10.3200/MONO.132.3.197-214

Hartwig, M., & Bond, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology, 28*(5), 661–676. https://doi.org/10.1002/acp.3052

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review, 19*(4), 307–342. https://doi.org/10.1177/1088868314556539

Hocking, J. E., & Leathers, D. G. (1980). Nonverbal indicators of deception: A new theoretical perspective. *Communication Monographs, 47*(2), 119–131. https://doi.org/10.1080/03637758009376025

Horvath, F. S. (1973). Verbal and nonverbal clues to truth and deception during polygraph examinations. *Journal of Police Science & Administration, 1*(2), 138–152.

Iacono, W. G., & Ben-Shakhar, G. (2019). Current status of forensic lie detection with the comparison question technique: An update of the 2003 National Academy of Sciences report on polygraph testing. *Law and Human Behavior, 43*(1), 86-98. http://dx.doi.org/10.1037/lhb0000307

Kassin, S. M., & Fong, C. T. (1999). "I'm innocent!": Effects of training on judgments of truth and deception in the interrogation room. *Law and Human Behavior, 23*(5), 499–516. https://doi.org/10.1023/A:1022330011811

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*(1), 165–172. https://doi.org/10.1037/0033-2909.112.1.165

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Koch, T., Holtmann, J., Bohn, J., & Eid, M. (2020). Multitrait-multimethod analysis. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 3034-3056). Springer International Publishing. https://doi.org/10.1007/978-3-319-24612-3_1331

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kraut, R. E., & Poe, D. B. (1980). Behavioral roots of person perception: The deception judgments of customs inspectors and laymen. *Journal of Personality and Social Psychology, 39*(5), 784–798. https://doi.org/10.1037/0022-3514.39.5.784

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods, 7*(2), 228–244. https://doi.org/10.1037/1082-989X.7.2.228

Landström, S., Granhag, P. A., & Hartwig, M. (2005). Witnesses appearing live versus on video: Effects on observers' perception, veracity assessments and memory. *Applied Cognitive Psychology, 19*(7), 913–933. https://doi.org/10.1002/acp.1131

Leal, S., Vrij, A., Fisher, R. P., & van Hooff, H. (2008). The time of the crime: Cognitively induced tonic arousal suppression when lying in a free recall context. *Acta Psychologica, 129*(1), 1–7. https://doi.org/10.1016/j.actpsy.2008.03.015

Leal, S., Vrij, A., Mann, S., & Fisher, R. P. (2010). Detecting true and false opinions: The Devil's Advocate approach as a lie detection aid. *Acta Psychologica, 134*(3), 323–329. https://doi.org/10.1016/j.actpsy.2010.03.005

Lilienfeld, S. O., & Andrews, B. P. (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal population. *Journal of Personality Assessment, 66*(3), 488–524. https://doi.org/10.1207/s15327752jpa6603_3

Luke, T. J. (2019). Lessons from Pinocchio: Cues to deception may be highly exaggerated. *Perspectives on Psychological Science, 14*(4), 646–671. https://doi.org/10.1177/1745691619838258

Lykken, D. T. (1974). Psychology and the lie detector industry. *American Psychologist, 29*(10), 725–739. https://doi.org/10.1037/h0037441

Lykken, D. T., & Venables, P. H. (1971). Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology, 8*(5), 656–672. https://doi.org/10.1111/j.1469-8986.1971.tb00501.x

Mann, S., & Vrij, A. (2006). Police officers' judgements of veracity, tenseness, cognitive load and attempted behavioural control in real-life police interviews. *Psychology, Crime & Law, 12*(3), 307–319. https://doi.org/10.1080/10683160600558444

Mann, S., Vrij, A., & Bull, R. (2002). Suspects, lies, and videotape: An analysis of authentic high-stake liars. *Law and Human Behavior, 26*(3), 365–376. https://doi.org/10.1023/A:1015332606792

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*(4), 335–361. https://doi.org/10.1177/014662168901300402

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00169

Melzack, R. (1975). The McGill Pain Questionnaire: Major properties and scoring methods. *Pain, 1*(3), 277–299. https://doi.org/10.1016/0304-3959(75)90044-5

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Paulhus, D. L. (1984). Two-component models of socially desirable responding. Jo*urnal of Personality and Social Psychology, 46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Picardi, C. A., & Masick, K. D. (2013). Reliability. In *Research methods: Designing and conducting research with a real-world focus* (pp. 43-53). Sage Publications, Inc.

Pennebaker, J. W. (1999). Psychological factors influencing the reporting of physical symptoms. In A. A. Stone, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (1st ed.) (pp. 299-315). Psychology Press. https://doi.org/10.4324/9781410601261

Podlesny, J. A., & Raskin, D. C. (1977). Physiological measures and the detection of deception. *Psychological Bulletin, 84*(4), 782–799. https://doi.org/10.1037/0033-2909.84.4.782

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Porter, S., & Brinke, L. (2010). The truth about lies: What works in detecting high-stakes deception? *Legal and Criminological Psychology, 15*(1), 57–75. https://doi.org/10.1348/135532509X433151

R. v. Béland, [1987] 2 SCR 398.

R v. Oickle, [2000] 2 SCR 3.

Rusbult, C. E., Martz, J. M., & Agnew, C. R. (1998). The Investment Model Scale: Measuring commitment level, satisfaction level, quality of alternatives, and investment size. *Personal Relationships, 5*(4), 357–391. https://doi.org/10.1111/j.1475-6811.1998.tb00177.x

Saxe, L., Dougherty, D., & Cross, T. (1985). The validity of polygraph testing: Scientific analysis and public controversy. *American Psychologist, 40*(3), 355–366. https://doi.org/10.1037/0003-066X.40.3.355

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*(5), 379–399. https://doi.org/10.1037/h0046234

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*(1), 1–22. https://doi.org/10.1177/014662168601000101

Spence, K., Villar, G., & Arciuli, J. (2012). Markers of deception in Italian speech. *Frontiers in Psychology, 3*. https://doi.org/10.3389/fpsyg.2012.00453

Sporer, S. L. (2016). Deception and cognitive load: Expanding our horizon with a working memory model. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00420

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology, 20*(4), 421–446. https://doi.org/10.1002/acp.1190

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*(1), 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Ströfer, S., Ufkes, E. G., Noordzij, M. L., & Giebels, E. (2016). Catching a deceiver in the act: Processes underlying deception in an interactive interview setting. *Applied Psychophysiology and Biofeedback, 41*(3), 349–362. https://doi.org/10.1007/s10484-016-9339-8

Strömwall, L. A., Hartwig, M., & Granhag, P. A. (2006). To act truthfully: Nonverbal behaviour and strategies during a police interrogation. *Psychology, Crime & Law, 12*(2), 207–219. https://doi.org/10.1080/10683160512331331328

Synnott, J., Dietzel, D., & Ioannou, M. (2015). A review of the polygraph: History, methodology and current status. *Crime Psychology Review, 1*(1), 59–83. https://doi.org/10.1080/23744006.2015.1060080

The Global Deception Research Team. (2006). A world of lies. *Journal of Cross-Cultural Psychology, 37*(1), 60–74. https://doi.org/10.1177/0022022105282295

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). John Wiley & Sons Ltd.

Vrij, A., Edward, K., & Bull, R. (2001). Police officers' ability to detect deceit: The benefit of indirect deception detection measures. *Legal and Criminological Psychology, 6*(2), 185–196. https://doi.org/10.1348/135532501168271

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior, 24*(4), 239–263. https://doi.org/10.1023/A:1006610329284

Vrij, A., & Fisher, R. P. (2020). Unraveling the misconception about deception and nervous behavior. *Frontiers in Psychology, 11*, 1377. https://doi.org/10.3389/fpsyg.2020.01377

Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology, 22*(1), 1–21. https://doi.org/10.1111/lcrp.12088

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences, 10*(4), 141–142. https://doi.org/10.1016/j.tics.2006.02.003

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling, 5*(1–2), 39–43. https://doi.org/10.1002/jip.82

Vrij, A., Fisher, R. P., Mann, S., & Leal, S. (2010). Lie detection: Pitfalls and opportunities. In G. D. Lassiter & C. A. Meissner (Eds.), *Police interrogations and false confessions: Current research, practice, and policy recommendations.* (pp. 97–110). American Psychological Association. https://doi.org/10.1037/12085-006

Walczyk, J. J., Harris, L. L., Duck, T. K., & Mulay, D. (2014). A social-cognitive framework for understanding serious lies: Activation-Decision-Construction-Action Theory. *New Ideas in Psychology, 34,* 22–36. https://doi.org/10.1016/j.newideapsych.2014.03.001

Woolridge, L. R., Leach, A.-M., & Elliott, E. (2020, March). *Perceptions of interviewees' accents during deception detection*. Paper presented at the annual meeting of the American Psychology-Law Society, New Orleans, LA, United States.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, pp. 1–59). Academic Press. https://doi.org/10.1016/S0065-2601(08)60369-X

Appendix A

Self-Report Measure (Study A)

1. How **nervous** were you when you were answering the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all nervous Extremely nervous

2. How **excited** did you feel when you were answering the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all excited Extremely excited

3. How **guilty** did you feel when you were answering the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all guilty Extremely guilty

4. How **surprised** were you when you were answering the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all surprised Extremely surprised

5. How **ashamed** were you when you were answering the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all ashamed Extremely ashamed

6. How **afraid** were you when you were answering the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all afraid Extremely afraid

7. How **anxious** were you when you were answering the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all anxious Extremely anxious

8. How **negative** did you feel when you were answering the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all negative Extremely negative

9. How **emotional** were you when you were answering the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all emotional Extremely emotional

10. How difficult was it for you to **answer** the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all difficult Extremely difficult

11. How difficult was it for you to **understand** the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all difficult Extremely difficult

12. How hard did you have to **think** about your answers?

1 2 3 4 5 6 7 8 9

Not at all hard Extremely hard

13. How hard did you have to **pay attention** to your behaviours?

1 2 3 4 5 6 7 8 9

Not at all hard Extremely hard

14. How hard did you have to **plan** your answers?

1 2 3 4 5 6 7 8 9

Not at all hard Extremely hard

15. How hard was it to **remember** your answers?

1 2 3 4 5 6 7 8 9

Not at all hard Extremely hard

16. How hard did you have to pay attention to **the examiners' behaviours**?

1 2 3 4 5 6 7 8 9

Not at all hard Extremely hard

17. How long did you **think** when answering the examiners' questions?

1 2 3 4 5 6 7 8 9

Not at all long Extremely long

19. How **motivated** were you to convince the examiners that you were telling the truth?

1 2 3 4 5 6 7 8 9

Not at all motivated Extremely motivated

20. Do you think that the examiners believed you? YES NO

Why?

_____

21. What behaviours do you think that the examiners were looking for?

_____

22. Would you prefer to be interviewed in (circle one):

YOUR NATIVE (FIRST) LANGUAGE YOUR NON-NATIVE LANGUAGE(S)

23. Please circle all of the items which you actually remember seeing in the video. Please **be honest** and tell the truth**. Do not lie**.

Cell phone	Bomb	Blueprints
Phone bills	Photo album	Calculator
Coffee mug	Printer	Wires
Energy drink cans	Bag of chips	Map
Gun	Plant	Calendar
Pens and pencils	Books	Gum
Tools (e.g., screwdriver)	Scissors	Newspaper clippings
Guitar	Drawing	Headphones

Appendix B

Self-Report Measure (Studies B and C)

1. How **nervous** were you when you were answering the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all nervous Extremely nervous

2. How **excited** did you feel when you were answering the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all excited Extremely excited

3. How **guilty** did you feel when you were answering the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all guilty Extremely guilty

4. How **surprised** were you when you were answering the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all surprised Extremely surprised

5. How **ashamed** were you when you were answering the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all ashamed Extremely ashamed

6. How **afraid** were you when you were answering the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all afraid Extremely afraid

7. How **anxious** were you when you were answering the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all anxious Extremely anxious

8. How **negative** were you when you were answering the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all negative Extremely negative

9. How **emotional** were you when you were answering the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all emotional Extremely emotional

10. How difficult was it for you to **answer** the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all difficult Extremely difficult

11. How difficult was it for you to **understand** the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all difficult Extremely difficult

12. How hard did you have to **think** about your answers?

1 2 3 4 5 6 7 8 9 10

Not at all hard Extremely hard

13. How hard did you have to **pay attention** to your behaviours?

1 2 3 4 5 6 7 8 9 10

Not at all hard Extremely hard

14. How hard did you have to **plan** your answers?

1 2 3 4 5 6 7 8 9 10

Not at all hard Extremely hard

15. How hard did you have to **remember** your answers?

1 2 3 4 5 6 7 8 9 10

Not at all hard Extremely hard

16. How hard did you have to pay attention to **the interviewer's behaviours**?

1 2 3 4 5 6 7 8 9 10

Not at all hard Extremely hard

17. How long did you **think** when answering the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all long Extremely long

18. How **detailed** were your answers to the interviewer's questions?

1 2 3 4 5 6 7 8 9 10

Not at all detailed Extremely detailed

19. How **motivated** were you to convince the interviewer that you were telling the truth?

1 2 3 4 5 6 7 8 9 10

Not at all motivated Extremely motivated

20. Do you think that the interviewer believed you? YES NO

Why?

_____

21. What behaviours do you think that the interviewer was looking for?

_____

22. Would you prefer to be interviewed in (circle one):

YOUR NATIVE (FIRST) LANGUAGE YOUR NON-NATIVE LANGUAGE(S)

23. Please circle all of the items which you actually remember seeing in the video. Please **be honest** and tell the truth**. Do not lie**.

Cell phone	Bomb	Blueprints
Phone bills	Photo album	Calculator
Coffee mug	Printer	Wires
Energy drink cans	Bag of chips	Map
Gun	Plant	Calendar
Pens and pencils	Books	Gum
Tools (e.g., screwdriver)	Scissors	Newspaper clippings
Guitar	Drawing	Headphones

Appendix C

Trained Coders' Observations Coding Manual

For each video assigned to you, please use the following information to code the respective measure. Please indicate on each scale on the Google Form the number that best represents your judgment.

Thinking Hard

- Overall, how hard did s/he have to think to tell his/her story and answer the questions? Use verbal and non-verbal information.
- Thinking hard is generally associated with lying.

- Please indicate on a scale from "Did not think hard" to "Thought extremely hard" how hard the story-teller had to think. If you are unsure, please mark "Unsure" in the middle of the scale.

1	2	3	4	5	6	7	8	9
Did not think hard				Unsure				Thought extremely hard

Nervousness

- Overall to you, does s/he appear anxious or is s/he comfortable? Consider both vocal and behavioral cues.
- Tension is generally associated with deception.
    - e.g., Is his voice tense? Is his posture rigid? Does he appear to be uncomfortable or nervous in the situation? Is he fidgeting? Does he avoid eye contact?

- Please indicate on a scale from "Extremely relaxed/comfortable" to "Extremely tense/nervous" how nervous or tense the story-teller appeared. If you are unsure, please mark "Unsure" in the middle of the scale.

1	2	3	4	5	6	7	8	9
Extremely relaxed/ comfortable				Unsure				Extremely tense/ nervous

Motivation

- Overall, how motivated is s/he to convince the interviewers s/he is telling the truth? Use verbal and non-verbal information.

- Please indicate on a scale from "Not at all motivated" to "Extremely motivated" how motivated the story-teller was to convince the examiners s/he was telling the truth. If you are unsure, please mark "Unsure" in the middle of the scale.

1	2	3	4	5	6	7	8	9
Not at all motivated				Unsure				Extremely motivated

Excitement

- Overall, how excited does s/he appear to be when answering the examiners' questions? Use verbal and non-verbal information.

- Please indicate on a scale from "Not at all excited" to "Extremely excited" how excited the story-teller appeared to be when answering the examiners' questions. If you are unsure, please mark "Unsure" in the middle of the scale.

1	2	3	4	5	6	7	8	9
Not at all excited				Unsure				Extremely excited

Appendix D

Trained Coders' Observations Form (Google Forms)

For each video assigned to you, please refer to the coding manual and use the following scales to code the respective measure. Please indicate on each scale the number that best supports your judgment.

1. Email
2. Participant ID#
3. Thought hard (5 = "Unsure")

1	2	3	4	5	6	7	8	9
Did not think hard				Unsure				Thought extremely hard

4. Nervousness (tense? fidgeting?) (5 = "Unsure")

1	2	3	4	5	6	7	8	9
Extremely relaxed/ comfortable				Unsure				Extremely tense/ nervous

5. Motivation (5 = "Unsure")

1	2	3	4	5	6	7	8	9
Not at all motivated				Unsure				Extremely motivated

6. Excitement (5 = "Unsure")

1	2	3	4	5	6	7	8	9
Not at all excited				Unsure				Extremely excited

Appendix E

Response Latency Coding Manual

Response Latency

Response latency is the length of time (in milliseconds) between the end of the interviewer's question (onset) and the beginning of the speaker's answer (offset) (Vrij et al., 2000).

Onset: the moment you stop hearing the interviewer/interpreter (whichever is later)
Offset: the moment you start hearing the interviewee talking

Please use the stable release of Datavyu (v:1.3.7) and ONLY .mp4 video files.

*If there is **NOT** already a file in the "Datavyu Response Latency Files" folder or the "Datavyu Response Latency Files > **Plane**" folder on Google Drive for the video you will be coding:*

1. Download the template.
2. Make a copy of the template on your computer and rename it to the participant (video) number.
3. Open Datavyu. Click File > Open and open the renamed template file.
4. There should be two columns (one for each coder). Be sure to use the column associated with your coder title (i.e., coder 1/3 or coder 2/4) by clicking on it.
5. Download the video file from either the "Edited Videos > **.mp4** Files" or the "Trimmed Videos — R > **.mp4** Files" (participant/video numbers with an 'R' suffix) folder.
6. In the controller, click Add Data to open the video file. Beside 'Plugin:' drop down to JavaFX Video. Select 'Open.'
7. Before coding, please be sure to manually set the frame rate by double clicking on *steps per second* and writing the correct frame rate. Please press Enter and Datavyu will use your new frame rate. (Refer to the frame rate reference for what the frame rate is.)
8. Click Play (or 8) to start the video.
9. Click 'Enter' at the end of the first question, and 9 at the start of the participant's response. If you make an error in the onset, you can rewind, highlight the onset, then press 7 to set the new onset. If you make an error in the offset, you can rewind, highlight the offset, then press 9 to set the new offset. If a participant cuts the interviewer/interpreter off, the onset and offset will be identical (i.e., the latency will be 0).
10. Repeat from step 8 for each question/latency (regardless of interviewer). The latencies between the interviewer/interpreter speaking and the participant responding are the only latencies of interest.
11. Be sure to save frequently (Ctrl/Cmd S) to avoid losing your work.

12. Finally, upload the completed Datavyu file for the participant to the "Datavyu Response Latency Files" folder on Google Drive.
13. Repeat from step 2 for each new participant (if there isn't already a data file on the Google Drive for that participant; otherwise, refer to the next section).

*If there **IS** already a file in the "Datayyu Response Latency Files" folder (NOT the Plane subfolder) on Google Drive for the video you have been assigned:*

1. Download the participant Datayyu file.
2. Open Datavyu. Click File > Open and open the participant Datavyu file.
3. There should be two columns (one for each coder). If there are multiple columns, it is likely the video has already been coded by two coders and does not need to be coded again. Please check that there are two *completed* columns for response latency somewhere in that file before moving on. This issue is addressed in the next section (re: the Plane subfolder) of this document. Be sure to use the column associated with your coder title (i.e., coder 1/3 or coder 2/4) by clicking on it.
4. Hide the other coder's column by selecting their column, clicking Spreadsheet > Hide Selected Columns. This is to ensure that the other coder's coding does not bias your coding.
5. Download the video file from either the "Edited Videos > **.mp4** Files" or the "Trimmed Videos — R > **.mp4** Files" (participant/video numbers with an 'R' suffix) folder.
6. In the controller, click Add Data to open the video file. Beside 'Plugin:' drop down to JavaFX Video. Select 'Open.'
7. Before coding, please be sure to manually set the frame rate by double clicking on *steps per second* and writing the correct frame rate. Please press Enter and Datavyu will use your new frame rate. (Refer to the frame rate reference for what the frame rate is.)
8. Click Play (or 8) to start the video.
9. Click 'Enter' at the end of the first question, and 9 at the start of the participant's response. If you make an error in the onset, you can rewind, highlight the onset, then press 7 to set the new onset. If you make an error in the offset, you can rewind, highlight the offset, then press 9 to set the new offset. If a participant cuts the interviewer/interpreter off, the onset and offset will be identical (i.e., the latency will be 0).
10. Repeat from step 8 for each question/latency (regardless of interviewer). The latencies between the interviewer/interpreter speaking and the participant responding are the only latencies of interest.
11. Be sure to save frequently (Ctrl/Cmd S) to avoid losing your work.
12. **IMPORTANT** (*steps 12, 13, 14, and 16 only apply if you are the second person coding a video*). After saving the completed file, please export the file to a .csv file. Name the .csv file the participant (video) number.
13. Open the .csv file and add a column before the first column.

14. Rename the blank first column to *Participant_ID* (**exactly**; case matters). Enter the participant ID number in the second row of the first column (i.e., after the *Participant_ID* heading).
15. Upload the updated Datavyu file for the participant to the "Datavyu Response Latency Files" folder on Google Drive. There should be data for RL1 and RL2 in this file. **It is okay to overwrite the previous file, although please be sure that the file you are uploading has two columns of completed response latency data.**
16. Finally, upload the .csv file to the ".csv Files" subfolder.
17. Repeat from step 1 for each new participant (if there is already a data file on Google Drive for that participant; otherwise, refer to the previous section).

*If there **IS** already a file in the "Datavyu Response Latency Files > **Plane**" folder on Google Drive for the video you have been assigned:*

1. Please check to ensure that there is data for response latency from two coders within the existing file.
2. If there is data in both of the columns:
    a. Please do not code that video, check it off/highlight it on your list (whichever method you are using to keep track of which videos you have coded), and move on to the next video.
3. If there isn't data for one or both coder(s):
    a. Please let Ryan know.

Controller Help

- Use set onset (Enter) to set the end of the question.
- Use set offset (9) to set the start of the response.
- Use 1 and 3 to skip through the video one frame at a time.
- Use 4 and 6 to speed up the video.
- Use find (+) after clicking on a cell. This will take you to that point in the video.
- Use hide tracks (*) to hide the track (less clutter).

Useful Tips

- Ctrl + \ can be used to delete a cell.
- After clicking a cell, you can then click "Snap Region" to constrain the boundaries to only play that clip.
    - To undo, select "Clear Region."
- To zoom on a specific region, click the zoom icon. To undo, click again.