International Journal of Communication 12(2018), 2461–2479 1932–8036/20180005

Scientific Evidence and Cue Theories in Deception Research: Reconciling Findings From Meta-Analyses and Primary Experiments

TIMOTHY R. LEVINE
University of Alabama at Birmingham, USA

A widely held belief about human communication is that specific verbal and nonverbal behaviors signal deception. This belief is held as folk wisdom across many cultures. It is also often portrayed as accepted social scientific knowledge in academic works. Explanations for why specific behaviors signal deception fall under the umbrella label of "cue theories." This commentary essay reviews the extensive social scientific theory and research on the utility of deception cues for detecting deception. Oddly, conclusions from meta-analyses do not align with the findings of the primary studies that comprise the meta-analyses. The divergent conclusions from meta-analyses and primary studies challenge both the validity of cue-based lie detection and what counts as the critical unit of scientific evidence in research. The implications for social science theory and research are discussed. Suggestions for improved applied lie detection are also provided.

Keywords: lying, nonverbal communication, meta-analysis, significance testing

There is a booming industry in teaching people to detect lies by reading others' nonverbal behavior. Former Central Intelligence Agency agents explain how to detect deception in *Spy the Lie* (Houston, Floyd, & Carnicero, 2012). Janine Driver, "lie detection expert for the FBI, CIA, and ATF," explains the "revolutionary program to supercharge your inner lie detector and get to the truth" in *You Can't Lie to Me* (Driver & Van Aalst, 2012, quotes from book cover). Would-be lie detectors can learn the Reid Technique by taking seminars from John E. Reid and Associates, who teach lie detection and interrogation to thousands of law enforcement professionals in the United States each year. You can test your "Lie-Q" online at http://liespotting.com/liespotting-basics/quiz/ or visit http://www.humintell.com/ for various products and workshops. And you can watch Pamela Meyer's TED talk (http://www.ted.com/speakers/pamela_meyer). This is only a small sampling of the deception cue training industry. Apparently, learning how to better detect lies is only a few clicks and some number of dollars away. One just needs to learn the secrets of reading nonverbal cues.

Strong and clear empirical evidence suggests that most people believe that nonverbal deception cues exist. Belief in the utility of cues has been long documented in meta-analysis (Zuckerman, DePaulo, & Rosenthal, 1981; see Hartwig & Bond, 2011, for the latest meta-analytic evidence) and in primary research

Timothy R. Levine: levinet111@gmail.com
Date submitted: 2017–07–17

showing that beliefs about deception cues transcend culture, religion, geographic region, and language (Bond & The Global Deception Research Team, 2006). People everywhere, for example, believe that liars avoid gaze and act nervous. Thus, selling people on nonverbal lie detection is not difficult because the message rings true to the target audience. Marketing nonverbal lie detection is preaching to the metaphorical choir.

There is also no shortage of academic theory and research on deception cues. Several academic theories in psychology and communication specify that, at least under certain conditions, deception is detectable based on the observation of nonverbal behavior. These theories are both old and new, and they have generated vast amounts of supportive and contradictory empirical evidence. There are currently at least seven published meta-analyses of cue utility in distinguishing truths from lies and three additional meta-analyses of the efficacy of cue training in lie detection.

The purpose of this review is to integrate findings on cue utility in lie detection across meta-analyses to understand the big picture of research findings on deception cues. Looking across findings leads to some stunning and perplexing revelations about cue theories and research findings that are invisible at the level of the individual experiment, program of research, or even individual meta-analysis. These observations lead me to ask: What's up with deception cues? The answer, I think, is something very strange. Here I document a pattern of results that, at least on the surface, runs counter to the usual understanding and practice of social science and statistics. The conclusions hold important implications for the verisimilitude of several theories of deception, the application of research findings to applied deception detection, and the very practice of modern social scientific inference.

Deception Cues

I define deception cues as objective, observable, and particular behaviors that are statistically associated with (a) folk wisdom regarding what liars do, (b) people's truth-lie assessments, and/or (c) the act of lying as opposed to telling the truth. That is, deception cues are specific behaviors thought to, or used to, rightly or wrongly distinguish truths from lies. The present focus is mostly on cues that consistently and reliably differentiate between deceptive and honest communication.

It is worth emphasizing that the cues are, at best, probabilistically related to deception rather than necessarily or invariably linked to deception. Research describes statistical trends over large numbers of people across a variety of situations. The idea that cues are probabilistic is completely noncontroversial in the deception research community, but the implications of the probabilistic nature of cues are often not fully appreciated.

As an example of a probabilistic cue, vocal pitch appears to be a reliable cue to deception (DePaulo et al., 2003). On average, liars tend to have a more highly pitched voice when they are lying than the voices of people speaking honestly. This does not mean that people with higher voices lie more than people with deeper voices or that if someone's vocal pitch increases, they are beginning to tell a lie. But, on average, lies are slightly more highly pitched than honest statements, all else being equal.

It is common in the social sciences (and deception research is no exception) to focus on significant differences. A result that is deemed statistically significant is one where the extent of difference or statistical association is improbable (usually 5% or less), assuming no difference or association at all. Statistical significance refers to the probability of a finding conditional on the statistical nil–null hypothesis being true. This means that presuming that just one focal test is conducted (which is usually not the case in deception cue literature), there is a 5% or smaller chance that the finding would have obtained if there were really no difference or association between the cue and actual honesty. In terms of effect size (which will be discussed shortly), a result that is significant is unlikely to be a nondifference of exactly $d = 0.00$.

It is valuable to know that some finding is unlikely to be exactly zero, but such knowledge is limited, and it is incorrect to make the inferential leap from a finding being not zero to a finding being something important or useful. If a person has $1, she is not flat broke, but that dollar will not take her very far or buy her much. Having a few dollars and being a millionaire are very different economic situations even though both are conditions different from having zero money.

So if some finding is not zero, we want to know whether the finding is relatively large or small. Statistical significance alone is insufficient. To better understand differences, we need to understand effect sizes, which indicate the strength of statistical association. In the case of deception cues, effect sizes indicate how much difference there is in the observation of the cue between truths and lies. Obviously, the larger the effect size associated with a given cue, the more useful that cue.

Because cues are probabilistically related to deception, using cues as the basis for deciding that a specific person is lying about something is inherently error-ridden. Again, the idea that deception cues are probabilistic and not deterministic is completely noncontroversial in deception research. To my knowledge, no one argues otherwise. What are controversial, however, are (a) the extent to which deception cues can be useful despite the imprecision; (b) which cues, if any, are useful; and (c) the conditions and processes moderating and mediating a cue's utility.

In addition to the probabilistic nature of cues, it is important to remember that cues are not enacted in isolation, even though they are often treated this way in the literature. That is, cues are often discussed (and treated statistically) as if they are independent of one another (e.g., Hartwig & Bond, 2011, and other lens model approaches). But cues can be highly intercorrelated. Cues present as constellations of cues, and they are interpreted by naive observers as a package or gestalt impression (Levine et al., 2011). These constellations of cues that create impressions comprise demeanor. When discussing one cue or another, it is important to remember that cues do not travel alone but rather as groups of cues. People's impressions and judgments rest more on overall demeanor than on any specific cue.

Cue Theories

I introduce the idea of *cue theories* (also see Levine & McCornack, 2014) as an umbrella term that captures and integrates the basic ideas throughout much of the past and present theory and research on deception detection. The divide between cue theories and noncue theories is perhaps the most fundamental issue in social scientific deception theory.

The core logic of cue theories is summarized as follows:

1. Telling truths and telling lies are psychologically different. How truths and lies are different varies from theory to theory, but examples include felt emotional states (fear of detection, guilt about lying, duping delight), the amount of autonomic nervous system arousal, the degree of cognitive load or effort naturally involved in message production, the impact of externally imposed cognitive load, strategic efforts to appear honest, message planning, and willingness to be forthcoming.

2. The psychological states produced by deception, in turn, are behaviorally signaled by specific observable cues. That is, the psychological states specified above mediate and explain the relationship between truths–lies and cues.

3. Therefore, deception can be detected (albeit indirectly and probabilistically) through observation of the cues arising from the mediating psychological states associated with deception. This is possible through either passive observation of cues or additional prompting to amplify the mediator. Regardless, all cue theories specify that deception can be detected by the observation of cues under certain theoretically specified conditions (e.g., high-stakes lies, observations of long naturalist interaction, or by adding additional cognitive load).

Examples of cue theories include Paul Ekman's approach (Ekman, 2009; Ekman & Friesen, 1969), four-factor theory (Zuckerman et al., 1981), interpersonal deception theory (Buller & Burgoon, 1996), and Aldert Vrij's approach of increasing cognitive load to increase the likelihood that some cues would be displayed (Vrij & Granhag, 2012; Vrij, Granhag, & Porter, 2010). The psychological mediators specified by each cue theory are often different. For example, Ekman focuses primarily on felt emotions, Vrij focuses on cognitive effort, and interpersonal deception theory examines strategic and nonstrategic cognitive and affective processes. The advocates of the various cue theories are likely to vehemently object to being lumped together with rival cue theories. Researchers devoted to specific cue theories are typically critical of other cue theories. But all cue theories share the same fundamental logical structure and differ only in the specific mediating process or processes and the moderating conditions.

All prominent cue theories specify critical boundary conditions that moderate the association between honesty–deceit and the specified internal psychological mediating process. The specification of these moderators provides theory-consistent but counterfactual explanations for aberrant results. When the data are not as predicted, this can be attributed to a failure to meet necessary boundary conditions, and thus null findings can be interpreted as theory-consistent. For Ekman, lies stakes serve this function. The emotions triggered by the act of lying are felt when the stakes are high, and a failure to observe cues means that the stakes were too low. For interpersonal deception theory, lies of insufficient communication duration enacted in insufficiently interactive settings are discounted. Vrij and Granhag (2012) hold that additional prompting is required. There is typically a circularity to the reasoning. For example, if an anticipated cue difference is not observed, then the stakes were too low. Evidence of low stakes is provided by the failure

to observe the predicted differences. This circularity serves to protect the various cue theories from nonsupportive findings.

Not all deception theories, however, share cue theory logic. DePaulo's (1992) self-presentation approach is predicated on honest and deceptive self-presentations being (mostly) psychologically similar. Everyone wants to create and maintain a positive impression and to be seen positively by others. Information manipulation theory 2 (McCornack, Morrison, Paik, Wiser, & Zhu, 2014) views the message production processes for truths and deception as the same. Similarly, cues and the psychological processes that produce cues play no role in accurate lie detection according to truth-default theory (Levine, 2014b), where the motives that guide communication are the same for truths and lies and the path to improved deception detection is through communication content, understanding context, and persuasion rather than the observation of cues. Nevertheless, several important deception theories past and present specify deception cues as the path to lie detection, and cue theories have generated much research.

Meta-Analyses of Deception Cues

Research on the validity–utility of cues in distinguishing truths and lies is extensive, and this research has been summarized over the years in at least seven different meta-analyses. Four of these meta-analyses focus on the utility of specific cues. The first of these meta-analyses was published in 1981 (Zuckerman et al., 1981). It examined 19 different cues that had been studied anywhere between two and 16 times each and found that eight of the 19 cues showed statistically significant differences between truths and lies. In 2003, the utility of deception cues was again assessed (DePaulo et al., 2003). This meta-analysis examined 88 cues or impressions from 120 different samples. Individual cues had been studied anywhere from three to 49 times. Two additional meta-analyses were published in 2006 and 2007 by Sporer and Schwandt; however, these were substantially smaller in scale than the 2003 analysis due to more conservative inclusion criteria. In addition to meta-analyses of specific cues, meta-analyses have examined the link between cue utility and cue use (Hartwig & Bond, 2011), how cue findings have changed over time (Bond, Levine, & Hartwig, 2015), and the predictive utility of multiple cues (Hartwig & Bond, 2014).

The results of the four meta-analyses examining specific cues are summarized in Table 1. Cues are divided into three groups: cues that show consistent differences between truths and lies across meta-analyses, cues that are significant in one meta-analysis but not others, and cues where the meta-analyses all agree that there is little difference between truths and lies. Table 2 summarizes the cues from the 2003 DePaulo et al. meta-analysis that had been studied at least 10 times.

Table 1. Validity of Deception Cues in Signaling Deception

	Absolute value of the effect size *d*		
Cue	Zuckerman et al. (1981)	DePaulo et al. (2003)	Sporer and Schwandt (2006, 2007)
Consistent differences between truths and lies			
Pupil dilation	1.49*	0.39*	—
Pitch	2.26*	0.21*	0.18*
Mixed findings			
Adaptors	0.40*	0.01	0.04
Head nods	—	0.01	0.18*
Hand movements	—	0.00	0.38*
Foot and leg movements	0.06	0.09	0.13*
Response latency	0.13	0.02	0.21*
Illustrators	0.12	0.14*	0.03
Repetitions	—	0.21*	0.17
Shrugs	0.48*	0.04	—
Speech errors	0.23*	0.17	0.08
Consistent no-difference			
Eye contact/gaze	0.11	0.01	0.02
Blinks	0.61	0.07	0.01
Head movements	0.27	0.02	0.12
Smile	0.09	0.00	0.06
Posture shift	0.08	0.05	0.02
Response length	0.12	0.03	0.08
Speech rate	0.02	0.07	0.02
Filled pauses	—	0.00	0.08
Unfilled pauses	—	0.04	0.03

* $p < .05$.

Table 2. Associations Between Cues and Lying (Cues Studied at Least 10 Times).

Cue	Number of prior studies	Effect size (*d*)	Heterogeneous
Number of details	24	−0.30*	Yes
Verbal–vocal uncertainty	10	+0.30*	No
Nervous	16	+0.27*	Yes
Vocal tension	10	+0.26*	Yes
Vocal pitch	21	+0.21*	Yes
Fidgeting	14	+0.16*	Yes
Illustrators	16	−0.14*	No
Facial pleasantness	13	−0.12*	Yes
Foot and leg movements	28	−0.09	
Speech rate	23	+0.07	
Blinking	17	+0.07	
Nonverbal immediacy	11	−0.07	
Posture shifts	29	+0.05	
Response length	49	−0.03	
Self-references	12	−0.03	
Response latency	32	+0.02	
Head movements	14	−0.02	
Relaxed posture	13	−0.02	
Eye contact	32	+0.01	
Self-fidgeting	18	−0.01	
Head nods	16	+0.01	
Silent pauses	15	+0.01	
Hand movements	29	.00	
Smiling	27	.00	
Non-”ah” speech errors	17	.00	
Filled pauses	16	.00	

* $p < .05$.

Across the four meta-analyses, two cues have consistently been found to distinguish lies from truth. Liars exhibit higher vocal pitch and larger pupil dilation than honest communicators. There were huge effects for these cues in 1981, but as the research has progressed, the cumulative effects have diminished and are no longer large. Thus, taken at face value, the best scientific evidence to date suggests that vocal pitch and pupil dilation are small but statistically significant cues to deception.

Next, one set of cues is statistically significant in one meta-analysis but not in the others. It is hard to know what to make of these inconsistencies. Take, for example, hand movements. In 2003, the effect is $d = 0.00$. In 2007, the effect is reported as $d = 0.38$. The 2007 effect is not large, but it is bigger than most, and it was, in fact, the largest effect reported in that meta-analysis. It was also highly statistically significant at $p < .001$. It is hard to reconcile how the cumulative across-study evidence can point to both zero and

significantly not zero. We might guess that a lot of supportive evidence accumulated between 2003 and 2007 to increase the effect from 0.00 to 0.38. But a closer look reveals that this cannot be the case. The 2003 d = 0.00 effect is based on N = 951 subjects from k = 29 studies, while the 2007 d = 0.38 effect is based on N =308 subjects from just k = 5 prior studies. The same pattern appears for response latency; the smaller effect in 2003 is based on more evidence than the larger effect in 2006. It is not clear why the more recent meta-analysis would be based on less evidence than an older meta-analysis, but different inclusion criteria is plausible. Regardless of the cause, the inconsistencies point to a phenomenon I have observed: Cue findings are ephemeral! Cue effects are highly significant in one study only to vanish or even reverse in the next. And the trend is, the more evidence, the smaller the effect. This issue is addressed in more detail in the next section on the decline effect.

The cues in the third set produce a consistent lack of difference between truths and lies. Cues such as eye contact, smiling, posture shifts, and speech rate show no significance differences and small effects across meta-analyses. It is scientifically safe to conclude that these behaviors do not usefully signal honesty or deceit. Although modern statistics does not accept the literal null hypothesis of d = 0.00, we can have a high degree of confidence that the true population effect is near zero.

Table 2 lists all the cues from the DePaulo et al. (2003) analysis that had been studied at least 10 times. The criterion of 10 or more prior studies is arbitrary, but because cue findings are shifty, it seems unwise to place too much confidence in findings based on less data than that. Table 2 lists some statistically significant cue effects, but most cues are not significant (even with sample sizes cumulated across 10 or more studies), and those cues that are significant show small and heterogeneous effects. Heterogeneous effects are those that vary significantly from study to study.

A Decline Effect for Cues

As mentioned earlier, cue findings seem erratic. Consider the especially well-designed and innovative cue experiment by deTurck and Miller (1985), which used an unusually realistic and rigorous method for generating truths and lies. They also developed a way to control for nondeception arousal. Six cues emerged in their results that differentiated not only between truths and lies but between aroused truths and lies: adaptors, hand gestures, speech errors, pauses, response latencies, and message duration. Moreover, the effect sizes ranged from d = 0.72 to d = 1.18. Clearly, strong effect sizes were reported. But consider also the meta-analysis findings available at the time (Zuckerman et al., 1981). Cumulative findings across studies were uniformly weaker in the meta-analysis, and the response latency and duration findings did not hold up as significant in meta-analysis. So which set of findings should we believe?

When the 2003 DePaulo et al. meta-analysis was eventually published more than a decade later, its findings discredited both prior candidates. The trend was clear. Effect sizes for cues become smaller over time as evidence accumulates. Maybe the careful reader already noticed this trend in Table 1.

A popular press news story provided a label for what I was observing: the "decline effect" (Lehrer, 2010). The title of the article was "The Truth Wears Off," and the subtitle asked, "Is There Something Wrong With the Scientific Method?" The article recounted several documented cases of once-good findings that

became increasingly difficult to replicate over time. Decline effects appear in the research on deception cues. The Hartwig and Bond (2011) meta-analysis conveniently listed side by side the average effect sizes and the number of studies comprising the average effect. The correlation was significant and negative. The decline effect is real, and the effect shows up both cross-sectionally and longitudinally. The decline effect in deception cue research is graphed in Figure 1. The trend is clear: Over time, as more evidence accumulates and the sheer amount of data increases, cues become weaker.

Of course, this is not how things typically work in social science. Usually, as sample size increases, findings are more likely to be significant (because statistical power improves). But for deception cues, the opposite is true. Collecting more data makes effects less likely to be significant.

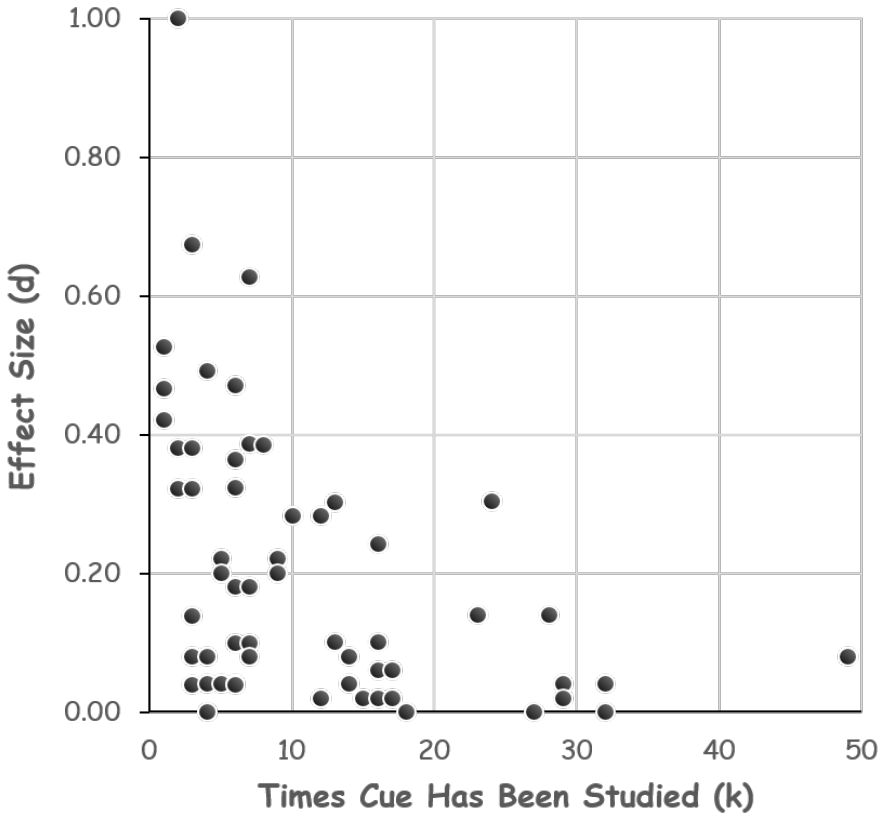

Figure 1. The relationship between the number of times a cue has been studied (k) and its cumulative effect size (d). Each data point is a cue in DePaulo et al. (2003). Thanks to Dr. Kim Serota for assistance in graph formatting.

Efficacy of Deception Detection Cue Training

If cue theories are valid, then training people to look for the most useful and reliable cues should improve deception detection. Various approaches have been used to train people to be better lie detectors, and three meta-analyses summarize the evidence for the efficacy of deception detection training (Driskell, 2012; Frank & Feeley, 2003; Hauch, Sporer, Michael, & Meissner, 2014). All three meta-analyses conclude that training significantly improves accuracy over no-training controls. In the 2003 analysis, accuracy in controls matched the 54% accuracy in the literature (Bond & DePaulo, 2006) compared with 58% for trained judges. The effect size for the improvement was d = 0.41. In the 2012 meta-analysis, the effect size was slightly larger (d = 0.50).

The evidence, however, might reasonably be construed as suggesting that training merely makes judges more cynical. Accuracy for detecting lies only improved from 49% to 55% with training, but accuracy for detecting truths dipped slightly from 58% to 56%. The most recent analysis, too, finds that gains are limited to accuracy for detecting lies; there was no improvement for truth accuracy (Hauch et al., 2014). Primary experiments have come to the same conclusion (Levine, Feeley, McCornack, Harms, & Hughes, 2005).

The evidence suggests that training improves accuracy by 4% or 5%, but there are four big caveats to consider. First, the improvement from training remains well within the slightly-better-than-chance range that is typical, with or without training (cf. Bond & DePaulo, 2006). This is the difference between getting 11 of 20 right and getting 12 of 20 correct. It is not particularly impressive. Second, as mentioned previously, training appears to affect cynicism more than ability, which is worrisome. Third, training in nonverbal cues is less effective than training that involves verbal content (Hauch et al., 2014). Finally, the gains are improvements over no-training controls. A more scientifically defensible design might involve a placebo control, and when testing against a placebo control, the efficacy of training diminishes substantially (Levine et al., 2005).

In 2005, a series of primary experiments compared three groups of judges: no training, valid training, and placebo training (Levine et al., 2005). Valid training and placebo training were created in two ways: either based on the research at the time or based on coding of the cues that were in the specific truths and lies that comprised the test materials. Valid training only improved accuracy when the training was based on cues present in and idiosyncratic to the materials used in the experiment. Training based on prior research findings did not improve accuracy. In the case where valid training was effective, it produced an 8-point gain in accuracy over the no-training control (50% to 58%) but only a marginal improvement over the placebo control (56% vs. 58%). Further, concerns about training producing skepticism were borne out. Both valid training and bogus training produced a greater proportion of lie judgments than the no-training controls (d = 0.81). Just as in the Frank and Feeley (2003) and Hauch et al. (2014) meta-analyses, training improvements emerged only for lies, not truths.

But Cues Are Potent in Individual Experiments

Based on the meta-analytic evidence reviewed above, one might be tempted to conclude that cues simply do not differ between truths and lies. The most recent meta-analytic evidence, however, strongly suggests otherwise. Hartwig and Bond (2014) examined the combined predictive power of multiple cues for distinguishing truths from lies in 125 prior research reports yielding 144 statistical predications. The findings are provocative and unexpected given the findings of previous meta-analyses. When the unit of analysis is the results of the individual deception cue experiment and not the specific cue under consideration, cues show substantial effects.

Hartwig and Bond's (2014) results are expressed in units of the multiple correlation coefficient (*R*), which reflects the combined predictive power of all the cues in each study. The average *R* was .52, a result that is highly significant and substantial in size. This translates to a percent correct rate of 67.9%. Clearly, cues do distinguish truths from lies quite well in the average study.

The extent of predictive power varied quite a bit from study to study. The multiple correlations ranged from near zero ($R = .01$) to almost perfect prediction ($R = .87$). The middle 50% of the findings fell between $R = .37$ and $R = .70$ with a median of $R = .48$. Despite the variation, the results were stable across several moderators. Student samples did not differ from nonstudent data. The motivation to lie (stakes) did not affect cue utility; nor did the presence or absence of interaction or whether the lies were about feelings or facts. Curiously, predictability was not affected by the number of cues to make the prediction or even the nature of the cues (visible, written, content, vocal, or global impression). Also surprising was that most of the predictive power was carried by a single cue ($r = .43$ for the strongest cue in each study, $R = .52$ for combined effects). Thus, the typical cue study provides strong evidence for at least one cue distinguishing between honest and deceptive communication. However, no supportive evidence was obtained for any of the moderators or mediators specified by any specific cue theory.

These findings may seem hard to reconcile with findings that focus on the utility of specific cues across studies. The $r = .43$ for the average effect for the strongest cue is equivalent to $d = 0.95$. As shown in Tables 1 and 2, however, there is no evidence that any specific cue is anywhere near this strongly associated with honesty and deceit. How can it be that some cue usually is associated with deception in published experiments, but no specific cues produce differences that can be replicated? How can $p < .001$ findings simply vanish or even reverse when tested again?

The findings are also hard to reconcile with the details of specific cue theories. Although the meta-analysis does document the predictive utility of cues at the level of the individual study, none of the key moderators (e.g., stakes, presences of interaction) specified by specific cue theories made any difference; nor were the cues associated with any specific mediator any more useful than any other mediator. No one explanation or boundary was any more viable than any other. All failed.

Perplexing Paradoxes

It is clear across several meta-analyses that specific nonverbal deception cues have little utility for deception detection. Summarizing the findings from the meta-analyses reviewed here, there is overwhelming scientific evidence that:

- At the level of the specific identifiable cue, most cues do not distinguish truths from lies at levels sufficient to rule out the statistical nil–null hypothesis.
- Of the few cues where the null hypothesis of no difference can be rejected, the effect sizes are small and heterogeneous and decline in magnitude as evidence accumulates.
- Training people in cues provides little practical gain in accuracy. Nonverbal lie detection training results in cynicism more than unbiased truth–lie discrimination and provides little gain over that of placebo effects.

Given all this evidence, how can we explain the continued scientific survival of cue theories and research? The results of meta-analyses appear to provide strong and compelling evidence that contradicts cue theory predictions and that discredits the utility of cue-based lie detection. Especially damaging is the inverse relationship between the amount of evidence and the supportiveness of the evidence. From a scientific perspective, the failure of supportive evidence to replicate and the large body of falsification evidence should lead to the demise of cue theories and research. But that has not happened.

The answer to the mystery of the survival of cue theories and research, I believe, lies in the ecology of modern social science research, where the important unit of evidence is the $p < .05$ finding in an original primary study and not the long-term replication and cross-validation of theoretically specified findings across studies and researchers. As the Hartwig and Bond (2014) meta-analysis reveals, at the level of the individual published article, substantial cue findings are the norm. In the primary studies examined by Hartwig and Bond, 136 of 144 (96%) found evidence of cue effects greater than or equal to $R = .20$. Consequently, authors of original research on deception cues can and do routinely obtain supportive findings from specific studies, and the advocates of various cue theories can and do cite long lists of citations in support of their views. Cue theorists and researchers can therefore selectively use cue meta-analysis as evidence against rival cues approaches while simultaneously claiming scientific evidence from primary studies for their own favored variety of cue theory. A clear example of this practice is Vrij and Granhag (2012), who use the DePaulo et al. (2003) meta-analysis to discredit Ekman, citing their own primary findings as evidence for their preferred view while ignoring the negative implications of the DePaulo meta-analysis findings for their own work. Simply put, when there are literally of hundreds of findings to choose from, it is easy to point to supportive evidence. At the same time, critics can point to findings supporting a completely different conclusion.

This account may also explain the persistence of folk beliefs about deception cues. Anecdotal evidence for cues is readily observed. Cues exist and are reified in personal experience. Just like individual researchers focused on their own results, everyday people in their social lives do not have access to the

bigger data perspective offered by meta-analysis. They experience the evidence for cues but not the disconfirming evidence. Belief in cues seems supported by experience.

This, then, is the real paradox. How can cues receive such strong support at the level of the primary study when that support so reliably evaporates at the level of meta-analysis of specific cues? A related question is whether social scientists need to rethink what counts as evidence or perhaps reject the scientific notions of falsifiability and replication as too burdensome and inconvenient. Maybe getting published and getting funded is more important than obtaining results that replicate across studies and different labs in modern social science.

Toward an Explanation

Some underappreciated findings might shed light on how some cues can be predictive of deceit in specific primary studies while specific cues are not consistently predictive across studies. Relevant here are Levine et al.'s (2005) placebo-control experiments testing the effectiveness of training people to detect deception with nonverbal cues. In their first study, the initial valid training condition used four cues that previous research suggested were associated with deception, and the placebo training condition involved four cues that past research had found unrelated to deception. Subjects were trained to look for either the valid cues, the placebo cues, or no cues at all (no-training control). The results were unexpected and initially inexplicable. The placebo group performed better than the no-training control, but the valid training group did the worst of all. The truths and lies were then coded for cues. In the truths and lies used in that first experiment, three of the four "valid" cues were significantly related to deception, but in the opposite way one would anticipate based on the literature. The cues were highly significant, but they flipped sign.

The stimulus messages used in the bogus training experiments involved two senders who produced four truths and four lies each. Four of eight cues showed significant differences for honesty, with effect sizes ranging from $d = 0.20$ to $d = 0.35$ (all wrong-direction inconsistent with meta-analysis findings). But these differences were small compared with the differences between the senders. The two senders differed on seven of eight cues ($d = 0.41$ to $d = 2.72$). Senders also differed in their cues from message to message within truths and with lies on four of the eight cues ($d = 0.41$ to $d = 0.84$). Further, statistically significant interactions occurred between sender, message, and honesty on 29 of the 32 possible interactions.

Consider the implications. There are cues differences at $p < .05$, but sometimes they flip sign. What signals deception in some studies signals honesty in others. Further, cues differ from person to person much more than they differ between truths and lies. And large differences are measured from message to message by the same person, holding honesty constant. People are not constant in the cues over time even within a situation, and that variation is *not* random.

Cues exist, and the statistical null is often false in any given data set. But cues are highly variable across message, sender, and context. The more we average across contingencies (i.e., higher order interactions), the more the average effect for cues regresses to zero. Significant differences in both directions average out. Interestingly, Kraut (1980) suggested this idea decades ago. What is different now is the existence of primary and meta-analytic data to document the ephemeral nature of cue findings.

The conclusion drawn from these findings is that compelling evidence exists for the ephemeral nature of cues. At the level of the individual study investigating some set of deception cues from some set of senders in some context, cues are statistically useful in distinguishing truths from lies. Cue studies reliably find cue effects that are statistically significant and that have moderate to large effect sizes, and this is true regardless of the demographics of the liar, the topic of the lies, the motivation of the liar, what the lie was about, or the cues studied (Hartwig & Bond, 2014). Yet when specific cues are studied again, the findings tend not to replicate. That is, there are usually cues that produce statistically significant results in most data sets, but what those cues tell us changes from study to study. Over time as research findings accumulate, the evidence for the utility of specific cues regresses toward zero. In short, cue findings seem to be informative only about the specific instances of communication under study and are not predictive or generalizable. The usual mode of statistical thinking based on sample size and the representativeness of the sample to the population falters. The important causal forces are instance-specific. Cue utility is idiosyncratic to a specific message by an individual communicator at a specific time in a certain situation.

Solutions

Is the inconsistency between the conclusions drawn from primary experiments and the conclusions from meta-analysis a problem? I argue that this question raises important meta-theoretical issues. The answer comes down to considerations of the desirability of using science-based criteria and what those scientific criteria should be.

Most deception theorists and researchers, I believe, want to claim scientific support for their views and their findings. Most deception research uses and relies on null hypothesis significance testing in original studies and experiments. Further, I suspect few deception researchers would publicly endorse selective use of evidence, the rejection of replicability as a standard, or see a lack of falsifiability as acceptable. Therefore, presuming desire for logical consistency, the state of affairs in cue research is clearly scientifically problematic, and the lack of a coherent empirical picture is disturbing. But this is only true if scientific standards are applied and if adherence to scientific standards includes the criteria that evidence cohere across studies.

I believe the problem largely exists because researchers knowingly or unknowingly exploit the statistical problem of overfitting and because cross-validation is seldom practiced in cue research. Overfit models are derived from and tested on idiosyncratic data and therefore do not replicate. Imagine cue study A looking at 10 cues coded from a sample of 100 honest and deceptive interviews. Logistic regression is used to predict honesty from the 10 cues, and a regression equation is obtained where each of 10 cues has an associated regression weight. Using those weights and cue scores, a multiple correlation is obtained, and percent correct classification can be calculated. The statistical software will pick optimal weights based on the data from study A. We know from meta-analysis that in data such as these, at least one cue is probably a substantial predictor and the correct classification is typically much better than chance. Then a different set of researchers perform cue study B using the same set of 10 cues to predict honesty–deceit in a different sample of interviews. We know from meta-analysis that, again, at least one cue is likely to be a substantial predictor and that the correct classification is likely to be substantially better than chance. But will it be the same cue or cues that are predictive in the same ways in both studies? Meta-analysis clearly says no. When

averaged across studies of the same cue, no cue has ever been found to produce strong and consistent findings. Thus, if we cross-validate by trying to predict honesty–deceit in study B's data with the regression equation used to successfully predict deception in study A, we would anticipate a substantial decline in predictive efficacy from one study to the next. The results of both studies A and B are overfit. The findings apply only to the truths and lies examined in the particular study. Within the confines of the individual study, the evidence looks strong. Those findings can be cited as evidence for the hypotheses and theory tested. The research can be published. But at some later time, when all the studies are meta-analyzed, that support vanishes, and the data show nonsupport after all. Consequently, the poor match between cue theories and data only begins to be apparent at the level of meta-analysis, and the true severity of the problems can be seen only when investigating several meta-analyses over time.

The first solution is obvious in principle but likely unpalatable in practice. Cue researchers could, as a matter of good practice, attempt to cross-validate their cue findings before they make them public. This involves building statistical models on one set of data and testing the model on other data sets. Supportive data would be more likely to replicate and would allow more confidence in the findings. A decade from now when meta-analyzed, supportive findings are likely to be far less ephemeral. The downsides, however, are that cross-validation is more effortful, and it is much less likely that the findings will support the preferred hypothesis or theory. It will be harder to publish research, and funders will question their support for one failed test after another. Cherished theories will be abandoned and there will be a search for hypotheses and theories that produce findings that replicate. This good-science approach works against the self-interests of too many researchers and labs.

An even more extreme and less attractive alternative is to employ statistical means of addressing the problem of overfitting. Researchers wanting to generalize beyond their data but not wanting to cross-validate the findings could treat senders, individual messages, and situations as random effects in their statistical models and adjust for the number of cues tested with Bonferroni-adjusted critical values. The problem with this approach is that it is so statistically underpowered that nonsignificance is almost guaranteed. The results, however, would likely be correct. Cue effects do not generalize, and this is just what Bonferroni-adjusted random-effects tests are likely to show. This approach is so underpowered that nothing would pass the test; thus, the cross-validation approach, although involving greater effort, is more promising.

Conclusion and Implications

For consumers interested in improved lie detection, whether an individual, a private business, a corporation, or a government, the implications are clear: Buyer beware. Although it is clear that specific falsehoods are often signaled by a specific nonverbal cue or sets of cues, it is also clear that no cue or set of cues is useful across instances, situations, and people. The use of nonverbal cues to detect deception guarantees errors—both false positives, where honest individuals are incorrectly identified as deceptive, and false negatives, were actual lies are missed. In a particularly telling recent field experiment, Ormerod and Dando (2014) tested two lie detection methods "in an in vivo double-blind randomized trial conducted in international airports" (p. 1). Security agents were given two weeks of training in either suspicious sign identification (nonverbal deception cues) or an alternative active content-based approach. The nonverbal

deception cue approach missed 97% of mock passengers passing through airport screening (i.e., the hit rate was just 3%). In contrast, the active content-based approach correctly identified two-thirds of the mock passengers. Readers are directed to Weinberger (2010) for a summary of the scientific research on using nonverbal cues to detect terrorists at airports and to Levine (2014a, 2015) for a review of evidence for the effectiveness of active, content-based lie detection. In sum, not only is nonverbal lie detection prone to error, but better alternatives are being developed and tested.

For deception theory, it is time to move beyond cue theories. The big problem for cue theories is that cumulative scientific data prove critical elements of every major cue theory false. The failure of supportive evidence to replicate across studies and the systematic trend for supportive evidence to become weaker as research progresses is highly damaging. Also damaging is the lack of evidence for critical moderators that various cue theories use to save theoretical predictions from the data. For example, in Ekman's program, deception stakes are a critical consideration, and cue utility is expected only for high-stakes lies. Nonsupportive findings are dismissed due to insufficient stakes. For the Vrij program, arousal-based cues are dismissed and prompted cognitive load cues are embraced. In interpersonal deception theory, only interactive deception of sufficient duration is expected to yield diagnostic cues, and short-duration or videotaped truths and lies are dismissed. Yet meta-analysis has examined these moderators, and cue utility is not moderated by consideration such as stakes, extent of interaction, or type of cues (Hartwig & Bond, 2014). Thus, not only is affirmative evidence lacking, but data undercut various cue theories' defense mechanisms. Further, as with the implications for applied deception detection, better alternatives now exist for theories, too.

The deception cue literature has important implications for how social scientists do social science. The conventional practice is to test hypotheses with null hypothesis significance tests and the almost exclusive reliance on *p* values generated from such tests. The deception cue literature provides a striking case study of how conventional practice can go wrong and thwart scientific progress. The negative association between amount of evidence and supportiveness of evidence is not unique to the deception literature and is, in fact, more common than not (Levine, Asada, & Carpenter, 2009). The limitations of significance testing are also well known (Levine, Weber, Hullett, Park, & Lindsey, 2008).

In conclusion, the scientific literature on the utility of nonverbal deception cues presents an apparent paradox. At the level of the individual experiment, ample evidence is obtained for the utility of cues in distinguishing honest from deceptive communication. Such findings, however, do not replicate. As evidence accumulates, the utility of specific cues diminishes. Cues are simply ephemeral. They are not random, and the support obtained in primary studies cannot be dismissed as Type I errors. Instead, variation in cues is highly contingent on micro variations in context, person, and time, which all interact in a way that prevents generalization across communication events. More robust approaches to data analysis are needed to document generalizable cue effects, and alternative theories and approaches are gaining currency.

References

Bond, C. F. Jr., & The Global Deception Research Team. (2006). A world of lies. *Journal of Cross-Cultural Psychology*, *37*, 60–74.

Bond, C. F. Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*, 214–234.

Bond, C. F. Jr., Levine, T. R., & Hartwig, M. (2015). New findings in nonverbal lie detection. In P. A. Granhag, A. Vrij, & B. Vershuere (Eds.), *Deception detection: Current challenges and new directions* (pp. 37–58). Chichester, UK: Wiley.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory*, *6*, 203–242.

DePaulo, B. M. (1992). Nonverbal behavior and self-presentation. *Psychological Bulletin*, *111*, 203–243.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*, 74–118.

deTurck, M. A., & Miller, G. R. (1985). Deception and arousal: Isolating the behavioral correlates of deception. *Human Communication Research*, *12*, 181–201.

Driskell, J. E. (2012) Effectiveness of deception detection training: A metaanalysis. *Psychology, Crime and Law*, *18*, 713–731.

Driver, J., & Van Aalst, M. (2012). *You can't lie to me*. New York, NY: Harper One.

Ekman, P. (2009). *Telling lies*. New York, NY: W. W. Norton.

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, *32*, 88–106.

Frank, M. G., & Feeley, T. H. (2003). To catch a liar: Challenges for research in lie detection training. *Journal of Applied Communication Research*, *31*, 58–75.

Hartwig, M., & Bond, C. F., Jr. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*, 643–659.

Hartwig, M., & Bond, C. F., Jr. (2014). Lie detection from multiple cues: Meta-analysis. *Applied Cognitive Psychology*, *28*(5), 661–676.

Hauch, V., Sporer, S. L., Michael, S. W., & Meissner, C. A. (2014). Does training improve the detection of deception: A meta-analysis. *Communication Research*, *43*, 283–343.

Houston, P., Floyd, M., & Carnicero, S. (2012). *Spy the lie*. New York, NY: St. Martin's Press.

Kraut, R. (1980). Humans as lie detectors: Some second thoughts. *Journal of Communication*, *30*, 209–218.

Lehrer, J. (2010, December 13). The truth wears off. *The New Yorker*. Retrieved from https://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off

Levine, T. R. (2014a). Active deception detection. *Policy Insights from the Behavioral and Brain Sciences*, *1*, 122–128.

Levine, T. R. (2014b). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*, 378–392.

Levine, T. R. (2015). New and improved accuracy findings in deception detection research. *Current Opinion in Psychology*, *6*, 1–5.

Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample size and effect size are negatively correlated in meta-analysis: Evidence and implications of a publication bias against non-significant findings. *Communication Monographs*, *76*, 286–302.

Levine, T. R., Feeley, T., McCornack, S. A., Harms, C., & Hughes, M. (2005). Testing the effects of nonverbal training on deception detection accuracy with the inclusion of a bogus train control group. *Western Journal of Communication*, *69*, 203–218.

Levine, T. R., & McCornack, S. A. (2014). Theorizing about deception. *Journal of Language and Social Psychology*, *33*, 431–440.

Levine, T. R., Serota, K. B. Shulman, H., Clare, D. D., Park, H. S., Shaw, A. S., Shim, J. C., & Lee, J. H. (2011). Sender demeanor: Individual differences in sender believability have a powerful impact on deception detection judgments. *Human Communication Research, 37*, 377–403.

Levine, T. R., Weber, R., Hullett, C. R., Park, H. S., & Lindsey, L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, *34*, 171–187.

McCornack, S. A., Morrison, K., Paik, J. E., Wiser, A. M., & Zhu, X. (2014). Information manipulation theory 2: A propositional theory of deceptive discourse production. *Journal of Language and Social Psychology*, *33*, 348–377.

Ormerod, T. C., & Dando, C. J. (2014). Finding a needle in a haystack: Toward a psychologically informed method for aviation security screening. *Journal of Experimental Psychology: General*, *144*(1), 76–84.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology*, *20*, 421–446.

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law*, *13*, 1–34.

Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, *1*, 110–117.

Vrij, A., Granhag, P. A., & Porter, S. B. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, *11*, 89–121.

Weinberger, S. (2010). Intent to deceive? Can the science of deception detection help catch terrorists? *Nature*, *465*, 412–415.

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). New York, NY: Academic Press.