MISUNDERSTANDINGS AND MISCONCEPTIONS IN THE APPLICATION OF NONVERBAL COMMUNICATION IN THE SPANISH LEGAL-FORENSIC CONTEXT

MALENTENDIDOS E IDEAS ERRÓNEAS EN LA APLICACIÓN DEL COMPORTAMIENTO NO VERBAL EN EL CONTEXTO JURÍDICO-FORENSE ESPAÑOL

ESTEBAN PUENTE-LÓPEZ[1], DAVID PINA[2], AND RAMÓN ARCE[3]

Cómo referenciar este artículo/How to reference this article:

Puente-López, E., Pina, D., & Arce, R. (2023). Misunderstandings and Misconceptions in the Application of Nonverbal Communication in the Spanish Legal-Forensic Context [Malentendidos e ideas erróneas en la aplicación del comportamiento no verbal en el contexto jurídico-forense español]. *Acción Psicológica, 20*(2), 43–70. https://doi.org/10.5944/ap.20.2.39334

Abstract

Background and objective: In both the general and specialized population, certain beliefs are held that relate nonverbal communication (NVC) to the probability that a person is telling the truth or lying. However, the evidence indicates that there is no nonverbal indicator or marker that can accurately and reliably discriminate between honest and dishonest testimony, warning of the pseudoscientificity of these techniques. Despite this, the use of NVC indicators is widely used in the fields of security and justice. Therefore, the present work aims to analyze the errors, practical utility and inadequate uses underlying the current practice of NVC in the Spanish legal-forensic context. **Method**: We made a review of the available evidence regarding NVC, as well as of the legal process and scientific criteria in the technical forensic field of the application of these techniques. **Conclusions**: The evidence on the use of NVC in the legal-forensic context is extremely limited, especially in the Spanish population, and does not meet the *Daubert* standards of admissibility of scientific evidence, i.e., judicially it is pseudoscientific evidence. This applies not only to its use in the detection of lying or deception, but also in all the practices in which the discipline has taken refuge, such as in processes of credibility of testimony in any type of crime, or in the assessment of emotional coherence or congruence.

Correspondence address [Dirección para correspondencia]: Department of Educational Sciences, Universidad de La Rioja, Logroño, Spain.
Email: david.pina@unirioja.es

ORCID: Esteban Puente-López (https://orcid.org/0000-0001-6367-457X), David Pina (https://orcid.org/0000-0001-5944-4683), and Ramón Arce (https://orcid.org/0000-0002-5622-3022).

[1] Universidad de Valladolid, Spain
[2] Universidad de La Rioja, Spain.
[3] Universidade de Santiago de Compostela, Spain.

Recibido: 20 de septiembre de 2023.

Keywords: Lie detection; Witness credibility; Daubert standards; Forensic assessment; Scientific evidence; Pseudoscientific evidence.

Resumen

Antecedentes y objetivos: Tanto en la población general como en la especializada se mantienen determinadas creencias que relacionan el Comportamiento No Verbal (CNV) con la probabilidad de que una persona esté diciendo la verdad o mintiendo. Sin embargo, la evidencia señala que no existe ningún indicador o marcador no verbal que permita discriminar con precisión y fiabilidad entre un testimonio honesto y deshonesto, advirtiendo de la pseudocientificidad de estas técnicas. A pesar de ello, el empleo de los indicadores de CNV está ampliamente extendido en los ámbitos de seguridad y justicia. Por ello, el presente trabajo tiene por objeto analizar los errores, utilidad práctica y usos inadecuados que subyacen en la práctica actual del CNV en el contexto jurídico-forense español. **Método**: Se ha realizado tanto una revisión de la evidencia disponible en lo relativo al CNV, como del proceso legal y criterios científicos en el ámbito técnico forense de la aplicación de estas técnicas. **Conclusiones**: La evidencia respecto al uso de CNV en el contexto jurídico-forense, especialmente en la población española, es notablemente limitada. Además, no cumple con los estándares Daubert de admisibilidad de la prueba científica, lo que judicialmente la califica como pseudocientífica. Esto no solo se aplica a su uso en la detección de la mentira o el engaño, sino también en todas las prácticas en las que se ha refugiado la disciplina, como en procesos de credibilidad del testimonio en cualquier tipología de delito o en la valoración de la coherencia o congruencia emocional.

Palabras clave: Detección de mentiras; Credibilidad del testimonio; Criterios Daubert; Evaluación forense; Prueba científica; Prueba pseudocientífica.

Introduction

Nonverbal communication (NVC) often described as language expressed through the face, body, or voice characteristics (Hall et al., 2019), i.e., any type of communication that excludes words (Denault et al., 2020). NVC has been the subject of interest and study for decades, both socially and scientifically, and as served as judicial evidence. The usefulness of the NVC lies in the belief that it offers the ability to identify and evaluate emotions, thoughts, and motives for people's behaviors (Schmid Mast & Hall, 2018). Such a belief has generated particular interest in contexts related to security, justice, and intelligence, given that it would allow determining, among other things, whether a person is lying (Denault et al., 2020; Patterson et al., 2023). Both in the general and specialized population (e.g., lawyers, judges, psychologists, criminologists, members of law enforcement agencies) this results in an (erroneous) assumption that body gestures or facial expressions can help determine whether a person is honest or dishonest (Vrij, 2008; The Global Deception Research Team, 2006).

Because of this (erroneous) assumption, many countries offer 'specialized' training courses that promise to teach justice and security professionals to 'accurately detect' lies using techniques based on scientific evidence. Some have even implemented NVC-based systems, such as the well-known Behavior Detection and Analysis (BDA) program, or the Screening of Passengers by Observation Techniques (SPOT), the latter used in airports in the United States of America to detect alleged terrorist threats based on the NVC and appearance of passengers (Brennen & Magnussen, 2020). In Spain, a quick search on the Internet is enough to identify multiple courses, experts, and masters. But none of them have an official status that, otherwise, would give graduates the official title qualifying them as expert witness referred to in papers 457-458 of the LeCrim and 340 of the LEC). The objective of these courses, experts and masters is usually to teach lie detection 'through rigorous and objective techniques from science', among others (e.g., Master in scientific non-verbal communication, human behavior and lie detection of the Fundeun).

Despite the apparent efficacy of NVC in catching liars preached by supposed experts and practitioners, decades of scientific evidence indicate that, to date, there is no nonverbal indicator or marker that can accurately and reliably discriminate between honest and dishonest testimony (Brennen & Magnussen, 2020; Vrij et al., 2019). In 2003, DePaulo et al. published the first meta-analysis 'Cues to Deception', with 1338 effect sizes of 138 NVC indicators associated with lying, finding that most indicators were not associated with deception and, if they were, the effect was very small. A few years later, Sporer and Schwandt (2006, 2007) replicated the NVC meta-analyses, divided into paraverbal and nonverbal cues associated with deception. Both studies identified a few cues reliably associated with deception (tone of voice, response latency and speech errors, assent, foot and leg movements, and hand movements), but in all cases the observed effects were extremely low, providing a discriminative ability close to chance, with opposite effects across studies and with opposite predictions depending on the theoretical model applied (i.e., if one theoretical model was applied it predicted an increase in the cue, but another or others predicted the opposite; idiosyncrasy error). For all these reasons, the authors concluded that these indicators and, by extension, the derived forensic techniques, lack scientific validity and, therefore, evidence in their own right, for the classification of false testimony. Vrij (2008; Vrij et al., 2019) warned that such indicators were unreliable and that their use in combination with other indicators was inadvisable as lie detection ability was diminished by focusing on them. Recently, a significant number of researchers signed a statement on questionable NVC-related practices used in security and justice contexts, warning of their pseudoscientific nature (Denault et al., 2020).

Despite all this, the use of NVC indicators in the fields of security and justice is still pervasive. In Spain, although there seems to be a relative acceptance that it is not a reliable tool for lie detection, its application has shifted to other equally questionable practices, such as the assessment of the emotional coherence of the victim-witness, or to support the processes of assessing the credibility of witnesses. Thus, a self-styled 'corps of behavior analysts' has been created, who apply various NVC techniques in the forensic field for various types of cases, such as gender violence, where the testimony of the alleged victim is often the only evidence available. These behavior analysts persist in the application of NVC in the forensic setting, likely because they fall into errors related to the area of psychology of testimony, methodology and data analysis, in the usefulness of the evaluation of testimony as evidence, not to mention the misconceptions about the uses of scientific evidence. For this reason, the current review aims to complete reports by Denault et al. (2020), Luke (2019) and Vrij et al. (2019) through the analysis of the errors, practical utility and inadequate uses underlying the current practice of NVC in the Spanish legal-forensic context.

What does it Mean to have Scientific Evidence?

In forensic psychology, evidence-based practice has been implemented with relative effectiveness, not only because of the availability of scientific knowledge, but also because it is in response to the judicial demand that experts provide evidence based on such scientific knowledge (art. 335.1 of the LEC). Arce (2017) outlined the criteria that a forensic technique has to meet to achieve in order to be qualified as scientific: (a) the measuring instrument must be reliable and valid; (b) the underlying technique must be falsifiable, refutable, replicable and testable; (c) the application of the technique must allow external review; (d) the methods used in the application of the technique must be verifiable; (e) the application of the technique to the case in question must be estimated; and (f) the technique must include an objective and strict decision criterion that fully controls false negatives - lying testimony classified as true - (criterial validity). Similarly, reference manuals in forensic psychology (Arce & Fariña, 2020; Carrasco-Ortiz & Rubio-Garay, 2020; Dujo-López et al., 2022; Sierra et al., 2010) place special emphasis on the fact that expert psychologists must base their reports on scientific evidence. The statement by Denault et al. (2020) also firmly establish this need in the field of NVC in particular.

In the last decade, professionals in the field of NVC have made an evident effort in trying to adapt to evidence-based practice. Paradoxically, many of these professional's advocate 'demystifying' the discipline, eliminating erroneous beliefs and educating on its scientific com-

ponent. They usually justify their practices with literature published in scientific journals, such as the meta-analyses of DePaulo et al. (2003) and claim that their practice has scientific rigor. Another quick search in the courses, experts and masters' web pages shows that they often state that they use techniques 'endorsed or validated by science' or 'from science'. Thus, it seems that one of the main misunderstandings that causes the use of NVC in the forensic context to persist in Spain is that professionals have a misconception of what is an evidence-based practice. This type of evidence seems to be understood as a dichotomous construct (either you have it or you don't) and appears to be used as a safeguard, or a symbol of quality, that justifies the use of their practices, when in fact, they are not ready for the contexts in which they are used. In other words, there is a simplistic view among supposed experts and practitioners that if a technique, or parts of it, has been subject to scientific studies, regardless of how they were conducted and where they are published, it is scientifically valid.

However, having scientific evidence does not unequivocally imply that a technique can be used in professional forensic practice. It is possible that, for methodological reasons, the quality of the evidence is poor, which affects the strength and validity of the inferences made, and drastically reduces its usefulness and admissibility in a judicial process (Gyuatt et al., 2008). It is also possible that, although the evidence is exemplary at the methodological level, it has used exclusively laboratory studies, and has not been tested in field studies (Arce, 2017; Puente-López et al., 2023). The existence of scientific evidence in itself should not be used as proof of sufficient quality of the technique, and the expert should not 'abdicate his/her responsibility as a scientist' in determining whether it can be used or not (Huss, 2014). For any test, tool, scale, technique, or methodology to be used, they must meet a series of scientific criteria or standards of quality that make them admissible before a judicial process (Rogers et al., 2023).

For several decades, what is known as *Daubert* standard (Daubert v. Merrell Dow Pharmaceuticals, 1993) have been used internationally to assess this issue. This standard explains the criteria evidence must meet to be admitted as scientific in courtrooms (Arce, 2017). The *Daubert* standard imposes the need to analyze in detail the quality of the evidence to be used and helps to partially solve the growing and severe problem of the use of pseudoscience or 'junk science' in the legal-forensic field. The *Daubert* standard adjusted to the case at hand would be: 1) Has the technique for evaluating testimony based on non-verbal behavior been tested? 2) Has the technique been peer reviewed and published? 3) Is the error rate of the technique known? 4) Are there guidelines for monitoring the performance of the technique? and 5) Does the technique enjoy general acceptance among the scientific community? We will now focus on criteria 1 to 3 since the meta-analysis reviewed do not provide scientific support for the indicators and technique, i.e., the consensus is that it is not a scientific test (criterion 5) and no guidelines (procedure) have been specified for the performance of the technique (criterion 4).

Daubert Criteria 1 and 2: Internal and External Validity.

Criteria 1 and 2 refer to two parts of what make up the process of generating scientific knowledge. The first criterion assesses whether the technique has been evaluated and tested through scientific research, and the second analyzes whether the knowledge generated in such research has been published in peer-reviewed journals. In this sense, it can be stated, with certain nuances, that the NVC technique in the legal-forensic context partially meets criterion 2, since a huge amount of scientific literature is available (Vrij et al., 2019). However, one may ask, to what extent is such evidence valid? Criterion 1 helps to answer this question with two fundamental concepts: internal and external validity.

In a scientific study, internal validity refers to the degree of certainty a researcher can have that the findings of his or her experiment are due to the manipulation of the variables selected for the study (Shadis et al., 2002). This implies that cofounding variables, which are those unmeasured variables that may act as an explanatory mechanism for the findings obtained in the study, have been controlled (Mehio-Sibai et al., 2004). On the other hand, external validity is the ability to extrapolate and generalize the results of a scientific study to different populations or situations (Mitchel, 2012). External validity is not

achieved through a single study, but through replication in different contexts and with different populations (McDermott, 2012). This is because in a single study, an estimate of validity in the population is obtained from a sample, so the error (sampling error) in the measurement is high. This is corrected as more samples are added (one sample always has less error than r; $\bar{r}$ always has less error than r). Although maximum external validity is desirable, it is obvious that in a lab a natural setting is impossible on many occasions, mainly for ethical reasons, such as in cases of child abuse, sexual assault, or gender violence. For this reason, when speaking of the external validity of studies in this context, the term face validity is used, and the aim is to recreate conditions that are as close as possible to the natural settings by means of research designs known as high-fidelity designs (Arce, 2017).

Thus, for an instrument, technique, or methodology to be administered in the legal-forensic context, scientific evidence is needed, but this evidence must have an adequate degree of internal and external/apparent validity, always taking into account the inherent limitations of the research topic. This means that the evidence cannot present methodological deficiencies that invalidate the conclusions reached, and it must be guaranteed that the findings have been obtained with a representative population, and randomly selected, of the one researcher want to evaluate. In this sense, Luke (2019) warned that the relevant literature on lie detection presented severe methodological deficiencies that limited the validity of their conclusions. For example, the system of NVC indicators commonly used to detect deception was derived by practitioners (never actually materialized into a technique as such) from the meta-analytic review of DePaulo et al. (2003). The derived system uses those NVC criteria with a significant effect on discriminating between those who lie and those who tell the truth from a total of 158 indicators, mostly NVC, but also verbal. However, after a reanalysis, Luke (2019) concluded that the effect of the NVC indicators often used by supposed experts and practitioners was not only small and unreliable, but that the average effect may be overestimated due to the low number of estimators and questionable methodological practices, such as selective reporting of information (publication bias) or the execution of studies with low statistical power.

Luke (2019) points out that, nowadays, the informative value of the evidence in the field of deception is very low, given that it is difficult to differentiate between real effects and false positives, and the state of the literature is compatible with the total absence of real indications of deception. To these limitations must be added the systematic use of convenience samples (Rosenthal effect), the lack of reliability in coding (coding reliability was not estimated) and, by extension, in the validity of the measurement of the indicators (it is not known exactly what they actually measured in each indicator, the unit of measurement is unknown, the definitions are not exhaustive so that the consistency in the inter-study measurement is unknown, nor totally independent -duplicity of measurements), and the high variability in the observed effect (average effect), such that the significant effects found may actually be trivial (in fact, DePaulo et al. take as significant effects if they are not zero, which does not really mean that the observed effect is significant). Although techniques generated by this means (indicators with a significant effect) is, in principle, exhaustive (the effects of all the indicators studied, a total of 138, were analyzed), the resulting system is not homogeneous or, at least, has not been contrasted. In sum, the system of indicators of NVC deception derived from the DePaulo et al. study does not meet the criterion of internal validity.

Another point to consider related to external/apparent validity is that none of the commonly recommended NVC indicators has been adapted to the Spanish population. Although the theory of the universality of emotions was assumed to be true for a few years, from 2010 onwards it began to be questioned, notably due to methodological flaws, and erroneous assumptions about the facial expression-emotion relationship (Patterson et al., 2023). Barrett et al. (2019), for example, showed that context can influence both the expressions produced and how they are perceived by others. For this reason, the evidence on the usefulness of NVC indicators when detecting lying obtained, for example, in the United States, the Netherlands, or the United Kingdom, cannot be uncritically generalizable to other cultural contexts such as Spain. Returning to the meta-analysis of DePaulo et al. (2003), a lack of criterion validity (significant negative effects) is also observed in part of the NVC indicators, i.e., the absence of the indicator is related to lying, the objective of the technique being

the detection of lying (the criterion cannot be supported by the absence of the indicator, but by its presence).

Later meta-analyses (Sporer & Schwandt, 2006, 2007) focused on specific signs of NVC (most of the indicators of DePaulo et al. disappeared from the list due to lack of productivity and lack of clarity in the definition, even some do not share the definition) which were divided into 9 paraverbal (Sporer & Schwandt, 2006) and 11 nonverbal (Sporer & Schwandt, 2007). Sporer and Schwandt found significant effects (results weighted by sampling error must be taken as unweighted results and be subject to a source of error that explains about 60% of the variance of the effects) in 2 paraverbals, tone of voice ($N = 284$, $r = .101$, 95% CI [-.017, .221]) and response latency ($N = 890$, $r = .106$, 95% CI [.037, .171]) and in 3 nonverbals, head nodding ($N = 590$, $r = -.091$, 95% CI[-.170, -.007]), foot and leg movements ($N = 799$, $r = -.067$, 95% CI[-.136, .006]), and hand movements ($N = 308$, $r = -.186$, 95% CI [-.293, -.073]). Reviewing the results, it is observed that in voice tone there is an error in the calculations since the effect is not really significant (the 95% confidence interval includes 0, with $Z = 1.27$, $p = .204$), as well as in foot and leg movements (the 95% confidence interval has 0, with $Z = -1.88$, $p = .060$). Thus, the significant effects are reduced to three indicators. Now, as already cautioned in the DePaulo et al. study, the Sporer and Schwandt define as a significant effect if it is not 0. It is obvious that an effect $\neq 0$ need not be a truly relevant effect. Thus, an r effect of .01 is not, de facto, significant. In this regard, Fandiño et al. (2021) defined that an effect was trivial (not significant) if $r \leq .05$ (the probability of classification error with the indicator would be $\geq .46$). Applying this criterion, the effect for the indicators response latency and head nodding would be trivial (the lower 95% CI limit is $< \pm.05$). Thus, only the effect of hand movements is significant (-.186). In conclusion, only 1 of the NVC indicators would have a truly significant effect size. In any case, the effect is negative, i.e., it does not detect deception, but identifies truth-tellers. Moreover, the predictions of the theoretical moldings are contradictory for this indicator. Thus, the results support (< hand movements in those who lie) the models of cognitive load (lying requires greater cognitive effort than telling the truth) and attempted control (liars attempt to control different communication channels to create credible behavior) but are contrary to the affective model (> hand movements in those who lie). In sum, scientific evidence does not validate the efficacy of NVC cues for deception detection.

Daubert Criterion 3: The Error Rate.

The fact that the technique or instrument has methodologically robust scientific evidence is not sufficient for it to be admissible in a judicial proceeding. It is not enough to know that the technique discriminates between groups with a given magnitude of effect; the degree of precision, and the probability of error in classifying a person as belonging to one of these groups must also be known. The third *Daubert* criterion implies that the error rate of the technique being used must be known and, in addition, this error rate must be appropriate for the domain of interest. In psychology, the difference between the observed result and the true result is known as measurement error. The error can result from various causes, such as the unreliability of the measurement instrument, the design or execution of the study, the subject's response style, or factors related to chance (VandenBos, 2015). Estimating the error rate is critical in the forensic context to determine the weight and performance of evidence (Dror & Scurich, 2020).

In forensic psychology, while 'error rate' is a broad concept in which several statistics for its estimation have a place, it has been commonly proposed to assess what are known as classification accuracy indices, which are sensitivity, specificity, and predictive power (Greve & Bianchini, 2004; Langleben & Moriarty, 2013). Following the definitions of Lange and Lippa (2017), sensitivity is the ratio of true positives of a test or measure, and it shows what percentage of the group of people who have a particular condition (e.g., are liars), have been correctly classified by the instrument. On the other hand, specificity is the ratio of true negatives of a test or measure and shows what percentage of the control group (without the condition studied, e.g., they are honest) have been correctly classified by the instrument used. Both indices can be expressed as a percentage, with a range from 0 to 100, or as a decimal fraction with a range from 0 to 1. By subtracting the value obtained from the maximum value (100 or 1), the false negative rate (identifying a liar as honest), in the case of

sensitivity, and false positive rate (identifying an honest person as a liar), in the case of specificity, can be obtained.

False positive and negative rates are, in this case, the values of interest in establishing the error rate of the technique, especially the false positive rate. For example, a test or technique to 'detect liars' with a specificity of 70% will have a false positive rate of 30%, implying that 3 out of 10 honest people will be misclassified as liars. In expert practice, the possibility of false positives (or negatives, when they imply the conviction of an innocent person as a result of the test) is unacceptable (Arce & Fariña, 2015). At all times, the expert's work must adhere to the constitutional principle of presumption of innocence and must avoid at all costs to incur in the erroneous classification of the person evaluated, since it would imply the conviction of an innocent person (Arce, 2017). Thus, it is more important not to commit a false positive than to commit a false negative (Sweet et al., 2021). The *Daubert* standard does not state what an 'acceptable' error rate is, and in forensic psychology, today, no consensus has been reached on a specific value. However, special emphasis has been placed on it being zero in theory (such an assumption must be supported, in the forensic setting, by a strict scientifically supported decision criterion). However, the probability of error rate of zero does not exist in measurement (Arce et al., 2010). In related disciplines, such as symptom simulation assessment, a false positive rate of between 5% and 10% was found to be acceptable for symptom and performance validity assessment tests that are commonly used in applied contexts where important decisions are made on the basis of test scores, such as forensics, since it generates an optimal trade-off between the two accuracy indices (Nunnally, 1978; Sweet et al., 2021).

Sensitivity and specificity help practitioners choose the most appropriate option for the task at hand, but do not provide information about the classificatory utility of a specific test or technique score in a particular individual (Smith et al., 2003). This issue can be addressed with test predictive values (Iverson, 2011; Lippa, 2018). Positive predictive value (PPV) refers to the proportion of people who are predicted by the test to have the condition and actually have it and negative predictive value (NPV) is concerned with the proportion of people who are predicted by the test not to have the condition and actually do not have it (Shreffler & Huecker, 2022). Thus, the predictive values allow us to answer the essential question of 'what is the probability that the person I am testing is a liar?' Both predictive values are calculated with the base rate, or prevalence, of the condition of interest in the particular context of study, in this case called the truth-lie rate (Iverson, 2011; Levine, 2018). A variation in such a rate will cause a variation in the classification accuracy of the technique or instrument, and if relatively certain values are not used the false positive rate may be underestimated (Levine et al., 2014; Shreffler & Huecker, 2022). For example, the false positive rate[3] of an instrument with 90% sensitivity and specificity will be much higher if the truth-lie rate in the context is 5% (PPV = 32%) than if it is 65% (PPV = 94%). For this reason, it cannot be assumed that the false positive rate identified in the United States, or in Finland, will be equally valid for Spain unless it is certain that the truth-lie rates are the same. Likewise, within the same country there will be variations in the truth-lie rate between contexts, such as between the judicial and administrative contexts in the evaluation of incapacity for work. Thus, it is critical that the interpretation of any scale, technique or instrument developed be limited to only those groups for which adequate data are available, especially when used in the context of high liability such as forensics (Rosenfeld et al., 2011).

Based on the above, the question arises: *do the NVC techniques meet the third Daubert criterion, do we know the error rate of the techniques used, and are these error rates admissible in the forensic context?* The answer is again no. Today, there is no peer-reviewed study available that evaluates the classification accuracy of any NVC technique in the Spanish forensic context, neither for detecting deception, nor for any of the other practices for which they are commonly used, such as the study of credibility or emotional congruence, which we will detail in the following section. In fact, precise estimates of the truth-lie rate in the different Spanish contexts of interest are not even available.

[3] In this case, the false positive rate would be calculated by subtracting the PPV value from 100%.

In the few international studies, such as those by Sporer and Schwandt (2006, 2007), or more recently by Matsumoto and Hwang (2018, 2020), an error rate of between 24% and 50% has been observed, which makes their use unfeasible in the forensic context. Now, as we reviewed above the results on the discriminative validity of the NVC deception indicators on which they base these estimates are not entirely correct. Be that as it may, the error rate of the forensic technique has never been tested, which, as we warned previously, has never been formulated as such. Even so, we can estimate this rate. On the one hand, the indicators that do not discriminate between honest and lying witnesses lack validity and are therefore not scientific, and only one indicator has been shown to be valid (r = -.186 for hand movements). From the observed effect size, we can calculate the success rate in correctly classifying the truth. Specifically, the probability of a witness being classified as a liar by this indicator is 64.8%, while the probability of being classified as honest is 35.2% (error not admissible in forensic evidence, punishable error).

Nonverbal Communication for Analyzing Credibility and Emotional Congruence.

The NVC technique in the Spanish forensic field has not been limited to lie detection and other alternative uses have been put forward: the analysis of credibility and emotional congruence. While it can be argued that both practices can be labeled as pseudoscientific, and that they do not meet any of the *Daubert criteria*, additional particularities should be highlighted.

Credibility of Testimony and NVC.

As Arce and Fariña (2006) point out, credibility can be understood from two approaches, one subjective, oriented to the social evaluation of the accuracy of the account, or apparent credibility of the witness, and the other objective, oriented to the evaluation of the accuracy of the account based on scientific evidence. In the forensic legal field, the analysis of the credibility of the testimony has become a fundamental pillar in most criminal cases that occur in the private sphere, such as sexual assault, gender violence or domestic violence (Novo & Seijo, 2010). In Spanish jurisprudence it is considered that the testimony of the victim-witness may be sufficient to overcome the presumption of innocence. However, it must have evidentiary aptitude for the judge to be certain about the veracity of the facts narrated by the complainant (Sentencia 210/2014 del TS[4]).

The presumption of innocence can only be rebutted when the statement of the complainant complies with a series of criteria that provide the necessary consistency to convey conviction about the facts. These are: absence of subjective incredibility, persistence in the incrimination and verisimilitude of the testimony (e.g., Sentencia 568/2016 del TS). The absence of subjective incredibility refers to the absence of any external incentive or motive (e.g., previous relationship between the complainant and the accused, resentment, enmity, revenge). On the other hand, persistence in the incrimination implies that the story must be coherent, without important modifications, persistent and concrete. Finally, the verisimilitude of the testimony implies that the testimony of the complainant must be supported by objective peripheral corroborations that provide the testimony with probative aptitude. This criterion is where the expert witnesses in the psychology of testimony comes in, and their work will be to apply a psychological test to endow the complainant's testimony with probative value and assist the judge/court in decision making on the credibility (the decision on credibility corresponds to the judge/court, not to the expert) of the person (Arce, 2017). As mentioned above, this expert evidence must have sufficient probative value to enervate the principle of presumption of innocence (no innocent person can be convicted; Sentencia 253/2004 del TS).

Although supposed experts and practitioners propose that the NVC can be used in the process of assessing the credibility of the testimony, it is not compatible, nor relevant, for this work. In practice, the NVC expert report will be administered to the victim-witness (although the technique is aimed at detecting lies, not to corroborate the truth of the witness), not on the testimony, and will evaluate

[4] "del TS" translates as "of the Spanish Supreme Court".

whether it appears to be credible (subjective or social credibility), a matter that is the sole and exclusive responsibility of the judge. Nevertheless, the expert evidence must provide the victim-witness's account with evidentiary aptitude (Arce, 2017). However, no scientific evidence of any kind is available on the relationship between NVC and victim-witness credibility. In fact, the scientific literature has been oriented to the relationship between NVC and lying, and scientific knowledge of lying has been extrapolated to credibility, equating two completely different constructs. Thus, a person can appear to be credible and not be honest, and vice versa (Köhnken et al., 2015). For this reason, indicators for lie detection are not valid, nor should they be used, to make any inferences regarding credibility. This is usually developed through what is known as verbal or statement content analysis (Vrij et al., 2019, Arce and Fariña, 2015), and involves identifying a series of reality criteria (SVA/CBCA; Steller and Köhnken, 1989), memory attributes (*Reality Monitoring*; Gancedo et al., 2021) or memory content criteria (SEG; Arce and Fariña, 2005) that indicate that the statement is proper to a real, perceived or self-experienced event (Vrij, 2015).

Emotional Congruence.

Another task that has been attributed to the NVC in the Spanish legal-forensic context is the determination of the emotional congruence/coherence of the witness-victim. For example, the Facial Expression Analysis Protocol (FEAP), commonly used in the expert practice of NVC in Spain, is presented as "a tool to detect the congruence or incongruence of the emotional facial expression (and not as) a tool to detect deception" (López-Pérez et al., 2016; p. 174). The technique proposes an analysis by levels where the expected emotion is compared with the emotion presented and it is proposed that, if the expected emotion coincides with the emotion shown, there will be "emotional congruence in the behavior developed (and) otherwise we should raise the hypothesis of emotional simulation" (p. 176).

This hypothesis in its application to the context of forensic assessment, and the premise of emotional coherence in general, rests on Paul Ekman's (1992) Theory of Basic Emotions. This theory assumes that there are basic emotions (anger, fear, happiness, sadness, and disgust) and universals that can be identified through facial expression (for an extensive critical review see Durán et al., 2017; Durán & Fernández-Dols, 2021). While these emotions are said to be reflected in the face, 'giving away' the actual emotion, a person may intentionally try to make their visible expressions discordant with what they actually feel (Crivelli & Fridlund, 2019). Within the framework of this theory, a hypothesis for witnesses (not testimony) evaluation is derived: if their facial expressions do not match their emotional state, the person is lying about that emotional state (Patterson et al., 2023). The full facial expression apparently reveals the emotion the person wants to show, but the authentic emotion can be leaked in the form of microexpression (leakage hypothesis, Ekman & Friesen, 1969; Vrij et al., 2019), an involuntary and fleeting emotional expression that lasts between 0.04 and 0.20 seconds (Matsumoto & Hwang, 2018).

The practice of this theory in the forensic context is highly inadvisable for the reasons already discussed in previous sections, but also because the concept of 'emotional congruence or coherence' is a particularly insidious premise. This premise is based on the existence of an 'emotional baseline' to be expected in each given situation, or for each type of victim. If, for example, the above-mentioned FEAP is administered to assess the emotional coherence of a victim of sexual assault or gender violence, it is assumed that evidence of the emotional profile of this victim typology will be available to establish what is to be expected and what is not. However, this approach is not supported by the scientific literature. Today there is no evidence, either nationally or internationally, of the emotional profile of any type of victim, in any type of context, and it is not possible to establish what is to be expected, beyond making inferences derived from a subjective, non-scientific, opinion. The expected emotions suggested in this type of practice are based on stereotypical reasoning about how the victim-witness should behave, which, in addition to having no scientific backing, may be totally erroneous and idiosyncratic (i.e., with contradictory interpretations). For example, in this type of report, the honesty of the smile may be analyzed, and highly questionable statements such as 'the alleged victim generates smiles inconsistent with sexual abuse' could be read. For the forensic context, such a statement requires answering to ques-

tions for which the scientific literature has no answer: How many incongruent smiles must occur to establish emotional incoherence? How does one combine the number of incongruent smiles with the rest of the incongruent emotions to establish an estimate of emotional incoherence? Does the presentation of the smile differ according to personal variables, diagnosis, traumatic event, etc.? What is the error rate associated with these indicators individually and overall? Moreover, as with micro expressions, the study of smiling has a questionable theoretical basis. De facto, it is based on the Duchenne smile hypothesis, according to which smiles that are the product of a genuinely positive emotion include eye constriction and allow us to differentiate when a smile is genuine or fake (Krumhuber & Manstead, 2009). Despite enjoying some acceptance, empirical support for this hypothesis has been mixed; it has been observed that they can occur when a person experiences negative emotions (Harris & Alvarado, 2005; Papa & Bonanno, 2008), that it is possible to fake them (Krumhuber & Manstead, 2009; Schmidt et al., 2009), and that, in general, rather than an indicator of a positive emotion it is more compatible with an artifact of an intense smile (Girard et al., 2021). However, even on the assumption that the Duchenne smile would make it possible to distinguish genuine from fake smiles, it must be necessary to answer another vital question: what does a genuine or fake smile imply in, for example, an alleged victim of gender-based violence? Thus, it is not possible to use this indicator, and any other, as evidence of 'emotional incongruence/inconsistency'. Any such inference totally lacks in scientific support and, therefore, should be inadmissible in a judicial proceeding.

Multiple Indicators and Instruments: Complementary Techniques and Protocols.

Another questionable idea is that NVC should be seen as a complementary tool to other techniques or instruments and used as a 'cross-cutting expert evidence', especially for testimony credibility cases. It is posited that NVC analysis is not the main element for decision making, but a technique that provides additional information in a multi-method strategy. Moreover, NVC supposed experts and practitioners urge not to look exclusively at those indicators that indicate 'lying', but at all possible indicators that can be obtained throughout the testimony. This approach is usually supported by what is commonly expressed in forensic psychology manuals (Carrasco-Ortiz & Rubio-Garay, 2020; Dujo-López et al., 2022; Sierra et al., 2010), that is, the importance of developing a psychological assessment tool based on multiple sources of information. Thus, it is common to find protocols, 'meta-protocols' or assessment systems that propose to combine multiple instruments or tools, generally content analysis techniques and NVC techniques (Vrij, 2015). Although those NVC proposals may seem appropriate, they have, in fact, more disadvantages than advantages.

When two or more techniques, instruments or tools are combined to assess the same variable, be it honesty, credibility, invalidity of symptoms, risk of recidivism, emotional consistency, etc., decision rules must be available to establish the status of the variable in question (Larrabe et al., 2019; Vrieze & Grove, 2010) and must increase the validity of the measure (Arce, 2017). No matter how organized and systematic protocols, 'meta-protocols' or assessment systems, may be, the formulation of hypotheses or conclusions that are not supported by statistically verified criteria are of no use to forensic practice, since there is no possibility of determining their degree of precision, and therefore, of error. Consider, for example, a presumed credibility assessment in which three components have been used: statement content analysis, facial analysis, and body analysis. What decision rule is followed to determine credibility? How are the results of such a combination evaluated? What should the practitioner do if the content analysis indicates sufficient credibility criteria, but the facial and body analysis results in “emotional inconsistency”? What accuracy and error rates are associated with each decision? Also, the addition of NVC techniques to others does not increase their validity as they have no scientific backing of efficacy in lie detection (and even less for the witness's honesty). On the contrary, it increases the noise (error rate).

Something similar occurs with the use of multiple NVC indicators. The literature reiterates that decision making should be derived from the combination of multiple nonverbal indicators (Hartwig et al., 2014; Matsumoto

& Wilson, 2023). While there are promising experimental studies on which NVC indicators to use together, and what estimated accuracy is derived from such joint use (Matsumoto & Hwang, 2018, 2020), robust scientific evidence on how to combine them effectively, and which decision rule to rely on to make a given conclusion or hypothesis, is lacking today. So, in a protocol combining, for example, the multiple expressive channels of NVC (facial expression, gestures, posture, orientation and movement, paralanguage, proxemics, etc.), what decision rule should you take for such a combination? How much data will be necessary to formulate a given hypothesis? What happens if discordant evidence is obtained? And more importantly, what proportion of indicators should be considered to estimate that there is a 'red flag'? The same conclusion should not be reached with 10% of suspected inconsistency indicators as with 50%, or 80%. Robust scientific evidence is lacking.

Another issue of great importance in the combination of multiple techniques or instruments is that evidence in favor of the use of individual elements should not be taken as evidence in favor of the use combinations of multiple techniques or instruments. In other words, if a component of a technique or instrument have 'scientific validity' by itself, it does not imply that the technique or instrument is validated. Common sense may indicate that the use of multiple instruments to assess the same variable will be a more robust and secure approach, but the truth is that this procedure increases the risk of false positives (in classifying a dishonest witness as honest, or false negatives in classifying an honest witness as dishonest) because the error ratio accumulates with the administration of each additional instrument (Bender & Frederick, 2018; Larrabe et al., 2019). Statistical analysis should explain what conditions the protocol must meet to measure the variable of interest satisfactorily. For example, in symptom simulation, this issue has been addressed in depth with symptom and performance validity tests (Erdodi, 2022; Larrabee, 2014; Larrabee, 2022; Larrabee et al., 2019). Although still a developing topic, some consensus has been reached that two positive results on two different symptom or performance validity tests (with a minimum of 90% specificity) in a protocol of up to 14 instruments, allow identification of an invalid presentation or low performance with an acceptable degree of certainty (Sweet et al., 2021).

Thus, the proposal to use techniques as complementary elements, or to combine techniques or instruments, without knowing their individual and overall performance, poses a severe operational problem. In the case of the NVC, no decision rules are available either for combining the elements that make up the technique (e.g., how to combine nonverbal indicators), or for combining it in conjunction with other instruments or techniques (e.g., with those of content analysis). Therefore, it is not possible to know their classification capacity or on its error rate, which makes it impossible them complementary techniques in the judicial process. Although it is possible to follow a decision system based on clinical judgment (clinical judgment is not scientific), it may be biased and it is impossible to provide a justification based on quantitative data of the decision taken (Dror & Scurich, 2020; Faust & Aherm, 2012; Garb, 2005). Furthermore, even if the status of the variable is not considered as binary (e.g., credible vs. not credible) and is proposed as a support in a hypothesis testing process, the error rate associated with the confirmation/rejection of each of the hypotheses must be known. The fact that the technique or tool is presented as a peripheral element whose result does not have to be evaluated in isolation should not be used to justify the introduction into the forensic assessment of practices whose impact, positive or negative, cannot be established with certainty.

Conclusions and Recommendations for Professional Practice.

Through the current review, it can be concluded that the evidence on the use of NVC in the legal-forensic context is extremely limited, especially in the Spanish population, and does not meet the *Daubert* standards of admissibility of scientific evidence, i.e., the use of NVC by supposed experts and practitioners amounts to pseudoscientific evidence. Scientifically, practically all the NVC indicators studied (DePaulo et al., 2003; Sporer & Schwandt, 2006, 2007) show a non-significant effect size. To reverse these results to a significant effect, a very large number of studies with significant effects on each of the NVC indicators would be necessary. And, since this is not possible, the conclusion about the lack of validity of the NVC indicators is definitive. Moreover, predictive models are contradictory: one model may predict one result for an indi-

cator and another just the opposite (idiosyncrasy error). This contradiction in prediction also unequivocally and definitively negates (lack of persistence in the judicial context, and of consistency in the scientific one) the validity of the measure.

In short, the use of NVC in forensic expert practice lacks judicial and scientific validity. This applies to all the practices in which NVC has taken refuge, whether in the detection of lies or deception, or in processes of credibility of testimony in any type of crime, its use to determine emotional coherence or congruence, or its use in general as a complementary or transversal technique to support other expert reports.

The application of NVC in the legal-forensic context, and security in general, is a young and developing discipline that has been given a professional applicability that goes beyond current evidence. The use of techniques that are in a premature and experimental state neglects the process of establishing robust theoretical frameworks and sound scientific practices on which to rely. Given the severe consequences of expert assessments in the legal-forensic context, selecting appropriate assessment tools is a fundamental responsibility of practitioners in the field (DeMatteo et al., 2019). Any conclusions reached by expert witnesses must be properly supported by the assessment system administered, and substantiated by quantitative evidence that allows the degree of precision of the estimates, as well as their error rate, to be assessed.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Ethical approval was not required.

Informed Consent Statement: Not applicable for studies not involving humans.

Data Availability Statement: The study does not report data.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

Arce, R. (2017). Análisis de contenido de las declaraciones de testigos: Evaluación de la validez científica y judicial de la hipótesis y la prueba [Content Analysis of Witness Statements: Assessing the Scientific and Judicial Validity of Hypothesis and Evidence]. *Acción Psicológica, 14*(2), 171–190. https://doi.org/10.5944/ap.14.1.21347

Arce, R. & Fariña, F. (2005). Peritación psicológica de la credibilidad del testimonio, la huella psíquica y la simulación: El Sistema de Evaluación Global (SEG) [Psychological Evidence in court on Statement Credibility, Psychological Injury and Malingering: The Global Evaluation System (GES)]. *Papeles del Psicólogo, 26,* 59–77. https://www.papelesdelpsicologo.es/pdf/1247.pdf

Arce, R. & Fariña, F. (2006). Psicología del testimonio y evaluación cognitiva de la veracidad de testimonios y declaraciones [Psychology of Testimony and Cognitive Assessment of the Veracity of Testimony and Statements]. In J. C. Sierra, E. M. Jiménez, & G. Buela-Casal (Eds.), *Psicología forense: Manual de técnicas y aplicaciones* (pp. 563–601). Biblioteca Nueva.

Arce, R. & Fariña, F. (2015). Evaluación psicológico-forense de la credibilidad y daño psíquico mediante el Sistema de Evaluación Global [Psychological-Forensic Assessment of Credibility and Psychic Damage using the Global Assessment System]. In P. Rivas & G. L. Barrios (Ed.), *Violencia de género: Perspectiva multidisciplinar y práctica forense* (pp. 411–441). Thomson Aranzadi.

Arce, R., Fariña, F., & Vilariño, M. (2010). Contrasting the Effectiveness of the CBCA in the Assessment of Credibility in Cases of Gender Violence. *Psychosocial Intervention, 19*(2), 109–119. https://doi.org/10.5093/in2010v19n2a2

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements. *Psychological*

Science in the Public Interest, 20(1), 1–68. https://doi.org/10.1177/1529100619832930

Brennen, T. & Magnussen, S. (2020). Research On Non-Verbal Signs of Lies and Deceit: A Blind Alley. *Frontiers in Psychology, 11*, Article 613410. https://doi.org/10.3389/fpsyg.2020.613410

Busch, R. & McCarthy, S. (2022). The Emergence of Evidence-Based Practice in Psychology. In J. N. Lester & M. O'Reilly (Eds.), *The Palgrave Encyclopedia of Critical Perspectives on Mental Health*. Palgrave Macmillan.

Carrasco-Ortiz, M. A. & Rubio-Garay, F. (Eds.). (2020). *Psicología jurídica y forense. Volumen I: Aspectos psicológicos y legales básicos* [Legal and Forensic Psychology. Volume I: Basic Psychological and Legal Aspects]. Sanz y Torres.

Crivelli, C. & Fridlund, A. J. (2019). Inside-out: From basic Emotions Theory to the Behavioral Ecology View. *Journal of Nonverbal Behavior, 43*, 161–194. https://doi.org/10.1007/s10919-019-00294-2

Daubert v. Merrell Dow Pharmaceuticals. (1993). 509 U.S. 579.

DeMatteo, D., Fishel, S., & Tansey, A. (2019). Expert Evidence: The (unfulfilled) Promise of Daubert. *Psychological Science in the Public Interest, 20*(3), 129–134. https://doi.org/10.1177/1529100619894336

Denault, V. (2020). Misconceptions about Nonverbal Cues to Deception: A Covert Threat to the Justice system? *Frontiers in Psychology, 11,* Article 573460. https://doi.org/10.3389/fpsyg.2020.573460

Denault, V., Plusquellec, P., Jupe, L. M., St-Yves, M., Dunbar, N. E., Hartwig, M., Sporer, S. L., Rioux-Turcotte, J., Jarry, J., Walsh, D., Otgaar, H., Viziteu, A., Talwar, V., Keatley, D. A., Blandón-Gitlin, I., Townson, C., Deslauriers-Varin, N., Lilienfeld, S. O., Patterson, M. L., ... van Koppen, P. J. (2020). The Analysis of Nonverbal Communication: The Dangers of Pseudoscience in Security and Justice Contexts. *Anuario de Psicología Jurídica, 30,* 1–12. https://doi.org/10.5093/apj2019a9

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

Dujo, V., González, D., & Graña, J. L. (2022). *Manual de psicología forense en el ámbito laboral* [Handbook of Forensic Psychology in the Workplace]. Pirámide.

Durán, J. I., & Fernández-Dols, J. M. (2021). Do Emotions Result in their Predicted Facial Expressions? A Meta-Analysis of Studies on the Link between Expression and Emotion. *Emotion, 21*(7), 1550–1569. https://doi.org/10.1037/emo0001015

Durán, J. I., Reisenzein, R., & Fernández-Dols, J. M. (2017). Coherence between Emotions and Facial Expressions: A Research Synthesis. In J. M. Fernández-Dols & J. A. Russell (Eds.), *The Science of Facial Expression* (pp. 107–129). Oxford University Press.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion, 6*(3-4), 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P. & Friesen, W. V. (1969). Nonverbal Leakage and Clues to Deception. *Psychiatry, 32*(1), 88-106. https://doi.org/10.1080/00332747.1969.11023575

Erdodi, L. A. (2023). Multivariate Models of Performance Validity: The Erdodi Index Captures the Dual Nature of Non-Credible Responding (continuous and categorical). *Assessment, 30*(5), 1467–1485. https://doi.org/10.1177/10731911221101910

Fandiño, R., Basanta, J., Sanmarco, J., Arce, R., & Fariña, F. (2021). Evaluation of the Executive Functioning and Psychological Adjustment of Child to Parent Offenders: Epidemiology and Quantification of Harm. *Frontiers in Psychology, 12*, Article 616855. https://doi.org/10.3389%2Ffpsyg.2021.616855

Faust, D. & Ahern, D. C. (2012). Clinical Judgment and Prediction. In D. Faust (Ed.), *Coping with Psychiatric and Psychological Testimony: Based on the Original Work by Jay Ziskin* (pp. 147–208). Oxford University Press.

Gancedo, Y., Fariña, F., Seijo, D., Vilariño, M., & Arce, R. (2021). Reality Monitoring: A Meta-analytical Review for Forensic Practice. *European Journal of Psychology Applied to Legal Context, 13*(2), 99–110. https://doi.org/10.5093/ejpalc2021a10

Garb H. N. (2005). Clinical Judgment and Decision Making. *Annual Review of Clinical Psychology,1,* 67–89. https://doi.org/10.1146/annurev.clinpsy.1.102803.143810

Greve, K. W. & Bianchini, K. J. (2004). Setting Empirical Cut-Offs on Psychometric Indicators of Negative Response Bias: A Methodological Commentary with Recommendations. *Archives of Clinical Neuropsychology, 19*(4), 533–541. https://doi.org/10.1016/j.acn.2003.08.002

Girard, J. M., Cohn, J. F., Yin, L., & Morency, L. P. (2021). Reconsidering the Duchenne Smile: Formalizing and Testing Hypotheses about Eye Constriction and Positive Emotion. *Affective Science, 2*(1), 32–47. https://doi.org/10.1007/s42761-020-00030-w

Hall, J. A., Horgan, T. G., & Murphy, N. A. (2019). Nonverbal Communication. *Annual Review of Psychology, 70*, 271–294. https://doi.org/10.1146/annurev-psych-010418-103145

Harris, C. R. & Alvarado, N. (2005). Facial Expressions, Smile Types, and Self-report during Humour, Tickle, and Pain. *Cognition and Emotion, 19*(5), 655–669. https://doi.org/10.1080/02699930441000472

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

Huss, M. T. (2014). *Forensic Psychology* (2nd ed.). Wiley.

Iverson, G.L. (2011). Positive Predictive Power. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology* (pp. 1968–1970). Springer. https://doi.org/10.1007/978-0-387-79948-3_1234

Köhnken, G., Manzanero, A. L., & Scott, M. T. (2015). Análisis de la validez de las declaraciones: Mitos y limitaciones [Statement Validity Analysis: Myths and Limitations]. *Anuario de Psicología Jurídica, 25*(1), 13–19. https://doi.org/10.1016/j.apj.2015.01.004

Krumhuber, E. G. & Manstead, A. S. (2009). Can Duchenne Smiles be Feigned? New Evidence on Felt and False Smiles. *Emotion, 9*(6), 807–820. https://doi.org/10.1037/a0017844

Lange, R. T. & Lippa, S. M. (2017). Sensitivity and Specificity should never be Interpreted in Isolation without Consideration of other Clinical Utility Metrics. *Clinical Neuropsychologist, 31*(6-7), 1015–1028. https://doi.org/10.1080/13854046.2017.1335438

Langleben, D. D. D. & Moriarty, J. C. (2013). Using Brain Imaging for Lie Detection: Where Science, Law, and Research Policy Collide. *Psychology, Public policy, and Law, 19*(2), 222–234. https://doi.org/10.1037/a0028841

Larrabee, G. J., Rohling, M. L., & Meyers, J. E. (2019). Use of Multiple Performance and Symptom Validity Measures: Determining the Optimal per Test Cutoff for Determination of Invalidity, Analysis of Skew, and Inter-test Correlations in Valid and Invalid Performance Groups. *Clinical Neuropsychologist, 33*(8), 1354–1372. https://doi.org/10.1080/13854046.2019.1614227

Larrabee, G. J. (2014). Aggregating across Multiple Indicators Improves the Detection of Malingering: Relationship to likelihood-ratios. *Clinical Neuropsychologist, 22*(4), 666–679. https://doi.org/10.1080/13854040701494987

Larrabee, G. J. (2022). Synthesizing Data to Reach Clinical Conclusion Regarding Validity Status. In R. W. Schroeder & P. K. Martin (Eds.), *Validity Assessment in Clinical Neuropsychological Practice; Evaluating and Managing Noncredible Performance* (pp. 193–210). The Guilford Press.

Levine, T. R. (2018). Ecological Validity and deception detection research design. *Communication Methods and Measures, 12*(1), 45–54. https://doi.org/10.1080/19312458.2017.1411471

Levine, T. R., Clare, D. D., Green, T., Serota, K. B., & Park, H. S. (2014). The Effects of Truth-Lie Base Rate on Interactive Deception Detection Accuracy. *Human Communication Research, 40*, 350–372. https://doi.org/10.1111/hcre.12027

Lippa S. M. (2018). Performance Validity Testing in Neuropsychology: A Clinical Guide, Critical Review, and Update on a rapidly Evolving Literature. *Clinical Neuropsychologist, 32*(3), 391–421. https://doi.org/10.1080/13854046.2017.1406146

López-Pérez, R., Gordillo, F., Soto, J. E., Pérez, M. A., & Salomoni, C. (2016). Protocolo FEAP (Facial Expression Analysis Protocol). In R. M. López, F. Gordillo, & M. Grau (Eds.), *Comportamiento no verbal* (pp. 179–192). Pirámide.

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

Matsumoto, D. & Hwang, H. C. (2018). Microexpressions Differentiate Truths from Lies about Future Malicious Intent. *Frontiers in Psychology, 9,* Article 2545. https://doi.org/10.3389/fpsyg.2018.02545

Matsumoto, D. & Hwang, H. C. (2020). Clusters of Nonverbal Behavior Differentiate Truths and Lies about Future Malicious Intent in Checkpoint Screening Interviews. *Psychiatry, Psychology, and Law, 28*(4), 463–478. https://doi.org/10.1080/13218719.2020.1794999

Matsumoto, D. & Wilson, M. (2023). Behavioral Indicators of Deception and Associated Mental States: Scientific Myths and Realities. *Journal of Nonverbal Behavior*. https://doi.org/10.1007/s10919-023-00441-w

McDermott, R. (2012). Internal and External Validity. In J. Druckman, D. Greene, J. Kuklinski, & A, Lupia (Eds.), *Cambridge Handbook of Experimental Political Science* (pp. 27–40). Cambridge University Press.

Mehio-Sibai, A., Feinleib, M., Sibai, T. A., & Armenian, H. K. (2004). A Positive or a Negative Confounding Variable? A Simple Teaching AID for Clinicians and Students. *Annals of Epidemiology, 15*(6), 421–423. https://doi.org/10.1016/j.annepidem.2004.10.004.

Mitchell G. (2012). Revisiting Truth or Triviality: The External Validity of Research in the psychological laboratory. *Perspectives on Psychological Science, 7*(2), 109–117. https://doi.org/10.1177/1745691611432343

Novo, M. & Seijo, D. (2010). Judicial Judgement-Making and legal Criteria of Testimonial Credibility. *European Journal of Psychology*

Applied to Legal Context, 2, 91–115. http://sepjf.webs.uvigo.es/index.php?option=com_docmanyx0026;task=doc_downloadyx0026;gid=26yx0026;Itemid=110yx0026;lang=en

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). McGraw-Hill.

Papa, A. & Bonanno, G. A. (2008). Smiling in the Face of Adversity: The Interpersonal and Intrapersonal Functions of Smiling. *Emotion, 8*(1), 1–12. https://doi.org/10.1037/1528-3542.8.1.1

Patterson, M. L., Fridlund, A. J., & Crivelli, C. (2023). Four Misconceptions about Nonverbal Communication. *Perspectives on Psychological Science, 18*(6), 1388–1411. https://doi.org/10.1177/17456916221148142

Puente-López, E., Pina, D., López-Nicolás, R., Iguacel, I., & Arce, R. (2023). The Inventory of Problems-29 (IOP-29): A Systematic Review and Bivariate Diagnostic Test Accuracy Meta-Analysis. *Psychological Assessment, 35*(4), 339–352. https://doi.org/10.1037/pas0001209

Rogers, R., Tazi, K. Y., & Drogin, E. Y. (2023). Forensic Assessment Instruments: Their Reliability and Applicability to Criminal Forensic Issues. *Behavioral Sciences & the Law, 41*(5) 1–17. https://doi.org/10.1002/bsl.2613

Rosenfeld, B., Edens, J., & Lowmaster, S. (2011). Measure Development in Forensic Psychology. In B. Rosenfeld & S. D. Penrod (Eds.), *Research methods in forensic psychology* (pp. 26–42). Wiley.

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

Sporer, S. L. & Schwandt, B. (2007). Moderators of Nonverbal Indicators of Deception: A Meta-analytic Synthesis. *Psychology, Public Policy, and Law, 13*(1), 1–34. https://doi.org/10.1037/1076-8971.13.1.1

Schmidt, K. L., Bhattacharya, S., & Denlinger, R. (2009). Comparison of Deliberate and Spontaneous Facial Movement in Smiles and Eyebrow Raises. *Journal of Nonverbal Behavior, 33*(1), 35–45. https://doi.org/10.1007/s10919-008-0058-6

Schmid Mast, M. & Hall, J. A. (2018). The Impact of Interpersonal Accuracy on Behavioral Outcomes. *Current Directions in Psychological Science, 27*(5), 309–314. https://doi.org/10.1177/0963721418758437

Sentencia 253/2004 del TS, Sala de lo Penal, de 04 de marzo de 2004. https://t.ly/ebpPd

Sentencia 210/2014 del TS, Sala de lo Penal, de 14 de marzo de 2014. https://vlex.es/vid/abuso-sexual-victima-declaraciones-503438218

Sentencia 568/2016 del TS, Sala de lo Penal, de 28 de junio de 2016. https://vlex.es/vid/644865189

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton, Mifflin and Company.

Shreffler, J. & Huecker, M. R. (2023). *Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values, and Likelihood Ratios*. StatPearls.

Sierra, J. C., Jiménez, E. V., & Buela-Casal, G. (2006). *Forensic Psychology: Manual of Techniques and Applications.* Biblioteca Nueva.

Smith, G. E., Cerham, J. H., & Ivnik, R. J. (2003). Diagnostic Validity. In D. S. Tulsky, D. H. Saklofske, G. J. Chelune, R. K. Heaton, R. J. Ivnik, R. Bornstein, A. Prifitera, & M. F. Ledbetter (Eds.), *Clinical Interpretation of the WAIS-III and WMS-III* (pp. 273-301). Academic Press.

Steller, M. & Köhnken, G. (1989). Criteria-Based Content Analysis. In D. C. Raskin (Ed.), *Psychological Methods in Criminal Investigation and Evidence* (pp. 217–245). Springer.

Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., Boone, K. B., Kirkwood, M. W., Schroeder, R. W., Suhr, J. A., & Conference Participants. (2021). American Academy of Clinical Neuropsychology (AACN) 2021 Consensus Statement on Validity Assessment: Update of the 2009 AACN Consensus Conference Statement on Neuropsychological Assessment of Effort, Response Bias, and Malingering. *Clinical Neuropsychologist, 35*(6), 1053–1106. https://doi.org/10.1080/13854046.2021.1896036

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

VandenBos, G. R. (Ed.). (2015). *APA Dictionary of Psychology* (2nd ed.). American Psychological Association.

Vrieze, S. I. & Grove, W. M. (2010). Multidimensional Assessment of Criminal Recidivism: Problems, Pitfalls, and Proposed Solutions. *Psychological Assessment, 22*(2), 382–395. https://doi.org/10.1037/a0019228

Vrij, A. (2008). *Detecting Lies and Deceit* (2nd ed.) Wiley.

Vrij, A. (2015). Verbal Lie Detection Tools: Statement Validity Analysis, Reality Monitoring, and Scientific Content Analysis. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Detecting Deception: Current Challenges and Cognitive Approaches* (pp. 3–35). Wiley-Blackwell.

Vrij A., Hartwig M., & Granhag P. A. (2019), Reading Lies: Non-verbal Communication and Deception. *Annual Review of Psychology, 70*, 294–317. https://doi.org/10.1146/annurev-psych-010418-103135

MALENTENDIDOS E IDEAS ERRÓNEAS EN LA APLICACIÓN DEL COMPORTAMIENTO NO VERBAL EN EL CONTEXTO JURÍDICO-FORENSE ESPAÑOL

ESTEBAN PUENTE-LÓPEZ, DAVID PINA Y RAMÓN ARCE

Introducción

Se conoce como comportamiento no verbal (CNV) al lenguaje expresado mediante la cara, el cuerpo o las características de la voz (Hall et al., 2019), es decir, cualquier tipo de comportamiento que excluya las palabras (Denault et al., 2020). El CNV ha sido objeto de interés y estudio desde hace décadas, tanto a nivel social como científico, y de valor como prueba judicial. La utilidad del CNV radica en que se le presupone la capacidad de identificar y evaluar las emociones, los pensamientos y los motivos de las conductas de las personas (Schmid Mast y Hall, 2018). Dicha capacidad ha generado especial interés en contextos relacionados con la seguridad, justicia e inteligencia, dado que permitiría determinar, entre otras cosas, si una persona está mintiendo (Denault et al., 2020; Patterson et al., 2023). Tanto en la población general como en la especializada (e.g., abogados, jueces, psicólogos, criminólogos, miembros de los cuerpos y fuerzas de seguridad del es-

tado) se mantienen determinadas creencias relacionadas con esta cuestión, como que los gestos corporales o las expresiones faciales pueden ayudar a determinar si una persona está diciendo la verdad o miente (Vrij, 2008; The Global Deception Research Team, 2006).

Tal es el interés que en múltiples países se ofrecen formaciones 'especializadas' que prometen enseñar a los profesionales de la justicia y seguridad a 'detectar con precisión' las mentiras mediante técnicas con evidencia científica. En algunos incluso se han instaurado sistemas basados en el CNV, como el conocido programa *Behavior Detection and Analysis* (BDA), antes conocido como *Screening of Passengers by Observation Techniques* (SPOT), utilizado en los aeropuertos de los Estados Unidos de América para detectar supuestas amenazas terroristas a partir del CNV y la apariencia de los pasajeros (Brennen y Magnussen, 2020). En España basta con una búsqueda rápida en internet para identificar múltiples cursos, expertos y másteres (ninguno con carácter oficial, lo que no confiere a los egresados el título oficial que habilita como perito titular al que se refieren los artículos 457-458 de la LeCrim y 340 de la LEC) cuyo objetivo es enseñar a detectar mentiras 'a través de técnicas rigurosas y objetivas procedentes de la ciencia', entre otros (e.g., Máster de comunicación no verbal científica, comportamiento humano y detección de mentiras de la Fundeun).

A pesar de la aparente eficacia del CNV para cazar a mentirosos predicada por supuestos expertos y profesionales, décadas de evidencia científica indican que, hoy en día, no existe ningún indicador o marcador no verbal que permita discriminar con precisión y fiabilidad entre un testimonio honesto y deshonesto (Brennen y Magnussen, 2020; Vrij et al., 2019). En el año 2003, DePaulo et al. publicaron el primer meta-análisis 'Cues to Deception', con 1338 tamaños del efecto de 138 indicadores de CNV asociados a la mentira, encontrando que la gran mayoría de los indicadores no se asociaban al engaño y, en el caso de hacerlo, el efecto era muy pequeño. Unos años después, Sporer y Schwandt (2006, 2007) replicaron los meta-análisis de CNV, divididos en indicios paraverbales y no verbales asociados al engaño. En ambos estudios se identificaron unos pocos indicadores asociados de forma fiable con el engaño (tono de voz, latencia de respuesta y errores de habla, asentimiento, movimientos de pies y piernas, y movimientos de manos), pero en todos los casos los efectos observados eran extremadamente bajos, proporcionando una capacidad discriminativa cercana al azar, con efectos contrarios entre estudios y con predicciones contrarias según el modelo teórico aplicado (es decir, si se aplicaba un modelo teórico se predecía un incremento en el indicador, pero otro u otros predecían lo contrario; error de idiosincrasia). Por todo ello, los autores concluyeron que estos indicadores y, por extensión, las técnicas forenses derivadas adolecen de validez científica y, por tanto, de prueba por sí mismos, para la clasificación de testimonios falsos. Vrij (2008; Vrij et al., 2019) advirtió que dichos indicadores eran poco fiables y que su uso en combinación con otros indicadores era desaconsejable ya que la capacidad de detección de la mentira disminuía al centrarse en ellos. Recientemente, un número importante de investigadores firmaron un comunicado sobre las prácticas cuestionables relacionadas con el CNV que se utilizaban en los contextos de seguridad y justicia, advirtiendo de la pseudocientificidad de las mismas (Denault et al., 2020).

A pesar de todo ello, el empleo de los indicadores de CNV en los ámbitos de seguridad y justicia dista mucho de estar en desuso o de un uso residual. En España, si bien parece haber relativa aceptación de que no es una herramienta fiable para la detección de la mentira, su aplicación ha virado a otras prácticas igual de cuestionables, como la valoración de la coherencia emocional de la víctima-testigo, o como apoyo en procesos de valoración de la credibilidad del testimonio. Así, se ha creado un autodenominado 'cuerpo de analistas del comportamiento', que aplican diversas técnicas de CNV en el ámbito forense para diversos tipos de casos, tal como la violencia de género, donde el testimonio de la presunta víctima suele ser la única evidencia disponible. Estos analistas persisten en la aplicación del CNV en el ámbito forense debido, principalmente, a que caen en errores relacionados con el área de la psicología del testimonio, metodología y análisis de datos, en la utilidad de la evaluación del testimonio como prueba, así como ideas equivocadas sobre los usos de la evidencia científica. Por este motivo, el presente trabajo tiene por objeto completar las lagunas de los informes de Denault et al. (2020), Luke (2019) y Vrij et al. (2019) a través del análisis de los errores, utilidad práctica y usos

inadecuados que subyacen en la práctica actual del CNV en el contexto jurídico-forense español.

¿Qué significa tener evidencia científica?

En psicología forense se ha implantado con relativa eficacia la práctica basada en la evidencia, no sólo por disponer de conocimiento científico, sino también porque es la respuesta a la demanda judicial a los peritos de aportar pruebas basadas en dichos conocimientos científicos (art. 335.1 de la LEC). Arce (2017) reseñó los criterios que ha de cumplir una técnica forense para alcanzar la característica de científica: a) El instrumento de medida ha de ser fiable y válido; b) la técnica subyacente debe ser falseable, refutable, replicable y sometible a prueba; c) la aplicación de la técnica ha de permitir la revisión externa; d) se han de poder comprobar los métodos usados en la aplicación de la técnica; e) se ha de estimar la aplicación de la técnica al caso en cuestión; y f) la técnica ha de incluir un criterio de decisión objetivo y estricto que controle totalmente los falsos negativos —testimonios mentirosos clasificados como verdaderos— (validez criterial). De la misma forma, los manuales de referencia en el área de psicología forense (Arce y Fariña, 2020; Carrasco-Ortiz y Rubio-Garay, 2020; Dujo-López et al., 2022; Sierra et al., 2010) enfatizan que los peritos psicólogos han de basar sus informes en evidencia científica. Además, el comunicado de Denault et al. (2020) estableció firmemente esta necesidad en el ámbito del CNV en particular.

En la última década, profesionales de dicho ámbito han hecho un evidente esfuerzo por tratar de adaptarse a la práctica basada en la evidencia. Paradójicamente, gran parte de estos profesionales abogan por 'desmitificar' la disciplina, eliminando creencias erróneas y educando en el componente científico de la misma. Habitualmente justifican sus prácticas con bibliografía publicada en revistas científicas, como los meta-análisis de DePaulo et al. (2003), y afirman que la práctica tiene rigor científico. Otra rápida búsqueda en los principales programas formativos y páginas de grupos de profesionales del CNV permite observar cómo en todo momento se afirma que se usan técnicas 'avaladas o validadas por la ciencia' o 'procedentes de la ciencia'. Así pues, parece que uno de los principales malentendidos que hace que el uso del CNV en el contexto forense persista en España es que los profesionales tienen una idea errónea de lo que es la práctica basada en la evidencia. Dicha evidencia parece ser entendida como un constructo dicotómico (se tiene o no se tiene) y se utiliza como un salvoconducto, o un símbolo de calidad, que justifica el uso de prácticas que no están preparadas para los contextos en las cuales se administran. Es decir, se plantea la visión simplista de que, si la técnica tiene estudios científicos, independientemente de cómo sean y dónde estén publicados, es científicamente válida.

Tener evidencia científica no implica inequívocamente que dicha evidencia pueda utilizarse en la práctica profesional forense. Cabe la posibilidad de que, por razones metodológicas, la calidad de la evidencia sea deficiente, lo que afecta a la fuerza y validez de las inferencias realizadas, y reduce drásticamente su utilidad y admisibilidad en un proceso judicial (Gyuatt et al., 2008). También es posible que, aunque la evidencia sea ejemplar a nivel metodológico, se hayan utilizado exclusivamente estudios de laboratorio, y no haya sido probada en estudios de campo (Arce, 2017; Puente-López et al., 2023). La existencia de evidencia científica en sí misma no debe utilizarse como prueba de suficiente calidad de la técnica, y el perito no debe 'abdicar de su responsabilidad como científico' a la hora de determinar su viabilidad (Huss, 2014). Un test, y cualquier prueba, herramienta, escala, técnica o metodología a utilizar, debe cumplir una serie de criterios o estándares científicos de calidad que hacen que sea admisible ante un proceso judicial (Rogers et al., 2023).

Durante varias décadas, se ha utilizado a nivel internacional para valorar esta cuestión lo que se conoce como criterios *Daubert* (Daubert v. Merrell Dow Pharmaceuticals, 1993), los cuales determinan el estándar que debe cumplir una prueba para ser admitida como científica en la Sala de Justicia (Arce, 2017). El estándar *Daubert* impone la necesidad de analizar pormenorizadamente la calidad de la prueba que se pretende utilizar, y ayuda a solventar, parcialmente, el creciente y severo problema del uso de pseudociencia o 'ciencia basura' en el ámbito jurídico-forense. Los criterios *Daubert* ajustados al caso que nos ocupa serían: 1) ¿Ha sido comprobada la técnica de evaluación del testimonio basada en el comportamiento no verbal?; 2) ¿Ha sido la técnica sometida a una revisión por

pares y publicada?; 3) ¿Se conoce la tasa de error de la técnica?; 4) ¿Existen directrices para el control de la ejecución de la técnica?; y 5) ¿Goza la técnica de una aceptación general entre la comunidad científica? Seguidamente nos centraremos en los criterios 1 al 3 dado que las revisiones meta-analíticas examinadas no prestan apoyo científico a los indicadores y técnica, es decir, el consenso es que no es una prueba científica (criterio 5) y no se han concretado directrices (procedimiento) para la ejecución de la técnica (criterio 4).

Criterios Daubert 1 y 2: Validez interna y externa

Los criterios 1 y 2 hacen referencia a dos de las partes que forman el proceso de generación de conocimiento científico. En el primer criterio se valora si la técnica ha sido evaluada y probada mediante una investigación científica, y la segunda analiza si el conocimiento generado en dichas investigaciones ha sido publicado en revistas que tengan un proceso de revisión por pares. En este sentido, puede afirmarse, con ciertos matices, que la técnica del CNV en el contexto jurídico-forense cumple parcialmente el criterio 2, ya que se dispone de una ingente cantidad de bibliografía científica (Vrij et al., 2019). No obstante, cabe preguntarse, ¿hasta qué punto dicha evidencia es válida? El criterio 1 ayuda a responder esta pregunta con dos conceptos fundamentales: la validez interna y externa. En un estudio científico, la validez interna hace referencia al grado de certeza que puede tener un investigador de que los hallazgos de su experimento se deben a la manipulación de las variables seleccionadas para el estudio (Shadis et al., 2002). Esto implica que se han controlado las variables extrañas, que son aquellas variables no medidas que pueden actuar como mecanismo explicativo de los hallazgos obtenidos en el estudio (Mehio-Sibai et al., 2004). Por otro lado, la validez externa es la capacidad de extrapolar y generalizar los resultados de un estudio científico a distintas poblaciones o situaciones (Mitchel, 2012). La validez externa no se alcanza mediante un solo estudio, sino a través de la replicación en diversos contextos y con diferentes poblaciones (McDermott, 2012). Esto es así porque de un único estudio se obtiene una estimación de la validez en la población a partir de una muestra, por lo que el error (error de muestreo) en la medida es elevado. Esto se corrige a medida que se añaden más muestras (una $\bar{r}$ siempre tiene menos error que *r)*. Si bien la máxima validez externa es deseable, es evidente que la recreación de las condiciones reales resulta imposible en muchas ocasiones, principalmente por razones éticas, como en casos de abusos a menores, agresiones sexuales, o violencia de género. Por este motivo, cuando se habla de la validez externa de los estudios en este contexto, se utiliza el término validez aparente, y se persigue recrear unas condiciones que sean lo más cercanas posibles a las reales mediante unos diseños de investigación conocidos como diseños de alta fidelidad (Arce, 2017).

Así, para que un instrumento, técnica o metodología se pueda administrar en el contexto jurídico-forense, no solo debe tener evidencia científica, si no que dicha evidencia debe presentar un grado adecuado de validez interna y externa/aparente, siempre teniendo en cuenta las limitaciones inherentes a la temática de investigación. Esto quiere decir que la evidencia no puede presentar deficiencias metodológicas que invaliden las conclusiones alcanzadas, y debe garantizarse que los hallazgos han sido obtenidos con una población representativa, y seleccionada al azar, de la que se pretende evaluar. En este sentido, Luke (2019) advirtió que la bibliografía relevante de la temática presentaba severas deficiencias metodológicas que limitaban la validez de sus conclusiones. Por ejemplo, el sistema de indicadores de CNV utilizado habitualmente para detectar el engaño fue derivado por los profesionales (en realidad nunca se materializó en una técnica como tal) de la revisión meta-analítica de DePaulo et al. (2003). El sistema derivado utiliza aquellos criterios de CNV con un efecto significativo en la discriminación entre los que mienten y los que dicen la verdad de un total de 158 indicadores, en su mayoría de CNV, pero también verbales. No obstante, tras una reanálisis, Luke (2019) concluyó que el efecto de los indicadores no sólo era pequeño y poco fiable, sino que el efecto promedio puede estar sobrestimado debido al bajo número de estimadores y a prácticas metodológicas cuestionables, como el reporte selectivo de información (sesgo de publicación) o la ejecución de estudios con baja potencia estadística. El autor señala que, hoy en día, el valor informativo de la evidencia en el ámbito del engaño es muy bajo, dado que resulta difícil diferenciar entre efectos reales y falsos positivos, y el estado de la bibliografía es compatible con la ausencia total de indicios reales de en-

gaño. A estas limitaciones hay que añadir el uso sistemático de muestras de conveniencia (efecto Rosenthal), la carencia de fiabilidad en la codificación (no se estimaba la fiabilidad de la codificación) y, por extensión, en la validez de la medida de los indicadores (no se sabe con exactitud qué han medido realmente en cada indicador, se desconoce la unidad de medida, las definiciones no son exhaustivas por lo que se desconoce la consistencia en la medida inter-estudios, ni totalmente independientes –duplicidad de medidas), y la alta variabilidad en el efecto observado (efecto promedio), tal que los efectos significativos hallados pueden ser realmente triviales (de hecho, DePaulo et al. toman como efectos significativos si no son cero, lo que realmente no quiere decir que el efecto observado sea significativo). Si bien, la técnica generada por este medio (indicadores con un efecto significativo), en principio, es exhaustiva (se analizaron los efectos de todos los indicadores estudiados, un total de 138), el sistema resultante no es homogéneo o, al menos, no se ha contrastado. En suma, el sistema de indicadores del engaño de CNV derivado del estudio de DePaulo et al. no cumple con el criterio de validez interna.

Otro punto por considerar relacionado con la validez externa/aparente, es que ninguno de los indicadores de CNV recomendados habitualmente ha sido validado en población española. Si bien la teoría de la universalidad de las emociones fue asumida como cierta durante unos años, a partir de 2010 comenzó a ponerse en duda debido a la limitada evidencia disponible y a planteamientos erróneos en la relación expresión facial-emoción (Patterson et al., 2023). Así, Barrett et al. (2019) mostraron que el contexto puede influir tanto en las expresiones producidas como en cómo son percibidas por parte de los demás. Por este motivo, la evidencia sobre la utilidad de indicadores de CNV a la hora de detectar la mentira obtenida, por ejemplo, en Estados Unidos, Holanda, o el Reino Unido, no son directamente generalizables a otros contextos culturales como España, debiendo pasar por un proceso de adaptación. Volviendo al meta-análisis de DePaulo et al. (2003), también se observa en parte de los indicadores de CNV falta de validez criterial (efectos negativos significativos), esto es, la ausencia del indicador se relaciona con la mentira, siendo el objetivo de la técnica la detección de la mentira (no se puede sustentar el criterio en la ausencia de éste, sino en la presencia).

En meta-análisis posteriores (Sporer y Schwandt, 2006, 2007) enfocados a los verdaderos indicios de CNV (la mayoría de los indicadores de DePaulo et al. desaparecieron del listado por falta de productividad y claridad en la definiciñon, incluso algunos no comparten la definición) que dividieron en nueve paraverbales (Sporer y Schwandt, 2006) y 11 no verbales (Sporer y Schwandt, 2007) hallaron efectos significativos (se han de tomar los resultados ponderados por el error de muestreo pues sin ponderar están sujetos a una fuente de error que explica en torno al 60% de la varianza de los efectos) en dos paraverbales, el tono de voz ($N = 284$, $r = .101$, IC 95% [-.017, .221]) y latencia de respuesta ($N = 890$, $r = .106$, IC 95% [.037, .171]) y en tres no verbales, asentir con la cabeza ($N = 590$, $r = -.091$, IC 95%[-.170, -.007]), movimientos de pies y piernas ($N = 799$, $r = -.067$, IC 95%[-.136, .006]), y movimientos de manos ($N = 308$, $r = -.186$, IC 95% [-.293, -.073]). Revisados los resultados se observa que en el tono de voz hay un error en los cálculos pues realmente no es significativo el efecto (el intervalo de confianza del 95% pasa por 0, siendo $Z = 1.27$, $p = .204$), al igual que en los movimientos de pies y piernas (el intervalo de confianza del 95% pasa por 0, con $Z = -1.88$, $p = .060$). Así pues, los efectos significativos se reducen a tres indicadores. Ahora bien, como ya se advirtió en el estudio de DePaulo et al., los autores definen como un efecto significativo si no es 0. Es obvio que un efecto $\neq 0$ no tiene por qué ser un efecto realmente relevante. Así, un efecto r de .01 no es, de facto, significativo. Al respecto, Fandiño et al. (2021) definieron que un efecto era trivial (no significativo) si $r \leq .05$ (la probabilidad de error en la clasificación con el indicador sería $\geq .46$). Aplicado este criterio, el efecto para los indicadores latencia de respuesta y asentir con la cabeza sería trivial (el límite inferior IC del 95% es $< \pm.05$). Así pues, sólo el efecto de los movimientos de manos es significativo (-.186). En conclusión, sólo uno de los indicadores del CNV tendría un tamaño del efecto realmente significativo. En todo caso, el efecto es negativo, es decir, no detecta engaño, sino que identifica a los que dicen la verdad. Además, las predicciones de los moldeos teóricos son contradictorias para este indicador. Así, los resultados apoyan (< movimientos de manos en los que mienten) los modelos de carga cognitiva (la mentira requiere un mayor esfuerzo cognitivo que la verdad) e intento de control (los mentirosos intentan controlar los indicadores de comportamiento no verbal para crear

una conducta creíble), pero son contarios al modelo afectivo (> movimientos de manos en los que mienten). En suma, la evidencia científica no valida la eficacia de los indicadores de CNV para la detección del engaño.

Criterio Daubert 3: La tasa de error

Que la técnica o instrumento disponga de evidencia científica metodológicamente robusta no es suficiente para que sea admisible en un proceso judicial. No basta con saber que dicha técnica discrimina entre grupos con una magnitud del efecto determinada, también debe conocerse el grado de precisión, y la probabilidad de errar, al clasificar a una persona como perteneciente a uno de dichos grupos. El tercer criterio *Daubert* comporta que debe conocerse la tasa de error de la técnica que se utiliza y, además, dicha tasa de error debe ser adecuada para el ámbito de interés. En psicología se conoce como error de medida a la diferencia entre el resultado observado y el resultado verdadero. El error puede ser consecuencia de diversas causas, como la falta de fiabilidad del instrumento de medida, del diseño o ejecución del estudio, del estilo de repuesta del sujeto o factores relacionados con el azar (VandenBos, 2015). La estimación de la tasa de error es fundamental en el contexto forense para determinar el peso y rendimiento de la evidencia (Dror y Scurich, 2020).

Si bien 'tasa de error' es un concepto amplio en el que tienen cabida varios estadísticos para su estimación, en psicología forense se ha propuesto habitualmente valorar lo que se conoce como índices de precisión de la clasificación que son la sensibilidad, la especificidad y el poder predictivo (Greve y Bianchini, 2004; Langleben y Moriarty, 2013). Siguiendo las definiciones de Lange y Lippa (2017), la sensibilidad es la ratio de auténticos positivos de una prueba o medida, y responde a la pregunta de qué porcentaje del grupo de personas que tienen una condición concreta (por ejemplo, son mentirosos), han sido correctamente clasificados por el instrumento. Por otro lado, la especificidad es la ratio de auténticos negativos de una prueba o medida, y responde a la pregunta de qué porcentaje del grupo control (sin la condición estudiada, por ejemplo, son honestos) han sido clasificados correctamente por el instrumento utilizado. Ambos índices se pueden expresar en porcentaje, con un rango del 0 a 100, o como fracción decimal con un rango del 0 al 1. Al restar el valor obtenido al valor máximo (100 o 1), se puede obtener la tasa de falsos negativos (identificar a un mentiroso como honesto), en el caso de la sensibilidad, y de falsos positivos (identificar a un honesto como mentiroso), en el caso de la especificidad.

Las tasas de falsos positivos y negativos son, en este caso, los valores de interés para establecer la tasa de error de la técnica, especialmente la tasa de falsos positivos. Por ejemplo, una prueba o técnica para 'detectar mentirosos' con una especificidad del 70 % tendrá una tasa de falsos positivos del 30 %, lo que implica que 3 de cada 10 personas honestas serán clasificadas erróneamente como mentirosas. En la práctica pericial la posibilidad de falsos positivos (o negativos, cuando impliquen derivado de la prueba la condena de un inocente) es inaceptable (Arce y Fariña, 2015). En todo momento la labor pericial ha de adherirse al principio constitucional de presunción de inocencia, y debe evitar a toda costa incurrir en la clasificación errónea de la persona evaluada, ya que implicaría en la condena de un inocente (Arce, 2017). Así, resulta más importante no cometer un falso positivo que cometer un falso negativo (Sweet et al., 2021). En el estándar Daubert no se establece qué es una tasa de error 'aceptable', y en psicología forense, hoy en día, no se ha alcanzado consenso sobre un valor concreto. No obstante, en teoría (dicho supuesto ha de estar avalado, en el ámbito forense, por un criterio de decisión estricto con amparo científico) ha de ser cero (aunque la probabilidad de error cero no existe en la medida; Arce et al., 2010). En disciplinas afines, como la evaluación de la simulación de síntomas, se determinó que una tasa de falsos positivos de entre el 5 % y el 10 % era aceptable para los test de evaluación de validez de síntomas y rendimiento que se utilizan habitualmente en contextos aplicados en los que se toman decisiones importantes a partir de las puntuaciones en los test, como el forense, dado que genera un intercambio óptimo entre los dos índices de precisión (Nunnally, 1978; Sweet et al., 2021).

La sensibilidad y la especificidad ayuda a los profesionales a elegir la opción más adecuada para la tarea en cuestión, pero no proporcionan información sobre la utilidad clasificatoria de la puntuación de una prueba o técnica específica en un individuo concreto (Smith et al., 2003). Esta cuestión puede abordarse con los valores predictivos de la

prueba (Iverson, 2011; Lippa, 2018). El Valor Predictivo Positivo (VPP) se refiere a la proporción de personas a las que la prueba predice que tienen la condición y realmente la tienen, y el Valor Predictivo Negativo (VPN) tiene por objeto la proporción de personas a las que la prueba predice que no tienen la condición y realmente no la tienen (Shreffler y Huecker, 2022). Así pues, los valores predictivos permiten responder a la pregunta esencial de '¿cuál es la probabilidad de que la persona que estoy evaluando sea un mentiroso?'. Ambos valores predictivos se calculan con la tasa base, o prevalencia, de la condición de interés en el contexto particular de estudio, en este caso llamada tasa verdad-mentira (Iverson, 2011; Levine, 2018). Una variación en dicha tasa provocará una variación en la precisión de clasificación de la técnica o instrumento, y si no se utilizan valores relativamente certeros puede subestimarse la tasa de falsos positivos (Levine et al., 2014; Shreffler y Huecker, 2022). Por ejemplo, la tasa de falsos positivos de un instrumento con 90 % de sensibilidad y especificidad será mucho mayor si la tasa de verdad-mentira en el contexto es del 5 % (PPV = 32 %) que si es del 65 % (PPV = 94 %). Por este motivo, no puede asumirse que la tasa de falsos positivos identificada en Estados Unidos, o en Finlandia, será igual de válida para España a no ser que se tenga la certeza de que las tasas verdad-mentira son las mismas. De la misma forma, dentro de un mismo país existirán variaciones en la tasa verdad-mentira entre contextos, como por ejemplo entre el judicial y el administrativo en las evaluaciones de la incapacidad laboral. Así pues, resulta crítico que la interpretación de cualquier escala, técnica o instrumento desarrollado se limite exclusivamente a aquellos grupos de los que se disponga datos adecuados, especialmente cuando se utilicen en el contexto de alta responsabilidad como el forense (Rosenfeld et al., 2011).

A partir de lo anteriormente expuesto cabe preguntarse: *¿cumplen las técnicas de CNV con el tercer criterio Daubert?, ¿conocemos la tasa de error de las técnicas utilizadas, y dichas tasas de error son admisibles en el contexto forense?* La respuesta vuelve a ser no. Hoy en día, no se dispone de ningún estudio revisado por pares que evalúe la precisión de la clasificación de ninguna técnica de CNV en el contexto forense español, ni para detectar el engaño, ni para ninguna de las otras prácticas para las que son utilizadas habitualmente, como el estudio de la credibilidad o de la congruencia emocional, las cuales detallaremos en el siguiente apartado. De hecho, ni si quiera se disponen de estimaciones precisas de la tasa verdad-mentira en los diferentes contextos españoles de interés.

En los pocos estudios internacionales, como los de Sporer y Schwandt (2006, 2007), o más recientemente los de Matsumoto y Hwang (2018, 2020), se ha observado una tasa de error de entre el 24 % y el 50 % (ver Vrij et al., 2019), lo que hace que su uso sea inviable en el contexto forense. Ahora bien, como revisamos anteriormente los resultados sobre la validez discriminativa de los indicadores de engaño del CNV sobre los que basan estas estimaciones no son totalmente correctos. Sea como fuere, nunca se ha puesto a prueba la tasa de error de la técnica forense, que, como ya advertimos previamente, nunca ha sido formulada como tal. Aun así, podemos estimar dicha tasa. Por una parte, los indicadores que no discriminan entre testigos honestos y mentirosos adolecen de validez y, por tanto, no son científicos, siendo sólo un indicador el que se ha mostrado válido (r = -.186 para los movimientos de manos). Del tamaño del efecto observado podemos calcular la tasa de acierto en la correcta clasificación de la verdad. En concreto, la probabilidad un testigo mentiroso sea clasificado como tal por este indicador es del 64.8 %, en tanto que la probabilidad de que sea clasificado como honesto es del 35.2 % (error no admisible en la prueba forense, error punible).

Comportamiento no verbal para analizar la credibilidad y la congruencia emocional

La técnica del CNV en el ámbito forense español no se ha limitado a la detección de la mentira y se han planteado otros usos alternativos: el análisis de la credibilidad y la congruencia emocional. Si bien ambas prácticas pueden tacharse directamente como pseudocientíficas, y puede afirmarse que no cumplen ninguno de los criterios *Daubert*, es necesario señalar una serie de particularidades.

Credibilidad del testimonio y CNV

Como señalan Arce y Fariña (2006) puede entenderse la credibilidad desde dos enfoques, uno subjetivo, orientado a la evaluación social de la exactitud del relato, o credibilidad aparente del testigo, y otro objetivo, orientado a la evaluación de la exactitud del relato basada en evidencia científica. En el ámbito jurídico-forense, el análisis de la credibilidad del testimonio se ha convertido en un pilar fundamental en la mayoría de los casos penales que ocurren en la esfera privada, tal como agresiones sexuales, violencia de género o violencia doméstica (Novo y Seijo, 2010). En la jurisprudencia española se considera que el testimonio del testigo-víctima puede ser suficiente para enervar la presunción de inocencia. No obstante, debe tener aptitud probatoria para que la persona juzgadora tenga certeza sobre la veracidad de los hechos narrados por la persona denunciante (Sentencia 210/2014 del TS).

La presunción de inocencia sólo puede enervarse cuando la declaración de la denunciante cumpla con una serie de criterios que otorga la consistencia necesaria para trasmitir convicción sobre los hechos. Estos son: ausencia de incredibilidad subjetiva, persistencia en la incriminación y verosimilitud del testimonio (e.g., Sentencia 568/2016 del TS). La ausencia de incredibilidad subjetiva hace referencia a que no concurra o concurriera en la denunciante algún tipo de incentivo externo o móvil (e.g., relación previa entre denunciante y acusado, resentimiento, enemistad, venganza). Por otro lado, la persistencia en la incriminación conlleva que el relato debe ser coherente, sin modificaciones importantes, persistente y concreto. Finalmente, la verosimilitud del testimonio implica que el testimonio del denunciante ha de estar avalado por corroboraciones periféricas objetivas que doten de aptitud probatoria al testimonio. Es en este criterio donde entra el perito experto en psicología del testimonio, y su labor será aplicar la prueba psicológica para dotar de valor probatorio al testimonio del denunciante, y asistir al juez/tribunal a tomar una decisión sobre la credibilidad (la decisión sobre la credibilidad corresponde al juez/tribunal, no al perito) de la persona (Arce, 2017). Y, recordemos, dicha prueba ha de tener valor probatorio suficiente para enervar el principio de presunción de inocencia (ningún inocente puede ser condenado; Sentencia 253/2004 del TS).

Aunque haya profesionales que plantean que el CNV puede utilizarse en el proceso de evaluación de la credibilidad del testimonio, no resulta compatible, ni pertinente, para esta labor. Procesalmente, la pericial de la CNV se administrará, para responder a la necesidad de dotar de valor de prueba al testimonio del denunciante, a la víctima-testigo (si bien, la técnica se orienta a la detección de la mentira, no para corroborar la verdad del testigo), no sobre el testimonio, y evaluará si aparenta ser creíble (credibilidad subjetiva o social), cuestión que compete única y exclusivamente al juez. Ahora bien, la labor del perito experto en los procesos de credibilidad del testimonio no ha de ser otra dotar de aptitud probatoria al relato de la víctima-testigo (Arce, 2017). Como consecuencia, no se dispone de evidencia científica de ningún tipo sobre la relación entre el CNV y la credibilidad del testigo-víctima. De hecho, la literatura científica se ha orientado a la relación entre el CNV y la mentira, y se ha extrapolado el conocimiento científico de la mentira a la credibilidad, equiparando dos constructos completamente diferentes. Así, una persona puede aparentar ser creíble y no ser honesto, y viceversa (Köhnken et al., 2015). Por este motivo, los indicadores para la detección de la mentira no son válidos, ni deben usarse, para hacer ninguna inferencia relativa a la credibilidad. Esto se desarrolla habitualmente mediante lo que se conoce como análisis verbal o del contenido de la declaración (Vrij et al., 2019, Arce y Fariña, 2015), e implica identificar una serie de criterios realidad (SVA/CBCA; Steller y Köhnken, 1989), atributos de memoria (*Reality Monitoring*; Gancedo et al., 2021) o criterios de contenido de la memoria (SEG; Arce y Fariña, 2005) que indiquen que la declaración es propia de un hecho real, percibido o autoexperimentado (Vrij, 2015).

Congruencia emocional

Otra de las tareas que se le ha atribuido al CNV en el contexto jurídico-forense español es el de la determinación de la congruencia/coherencia emocional del testigo-víctima. Por ejemplo, el *Protocolo de Análisis de la Expresión Facial* (Facial Expression Analysis Protocol, FEAP), utilizado habitualmente en la práctica pericial del CNV en España, se presenta como "una herramienta para detectar la congruencia o incongruencia de la expresión facial emocional (y no como) una herramienta para detec-

tar el engaño" (López-Pérez et al., 2016; p. 174). La técnica plantea un análisis por niveles donde se compara la emoción esperada con la emoción presentada y se propone que, si la emoción esperada coincide con la mostrada, existirá "congruencia emocional en la conducta desarrollada (y) en caso contrario deberíamos plantear la hipótesis de simulación emocional" (p. 176).

Esta hipótesis en su aplicación al contexto de evaluación forense, y la premisa de la coherencia emocional en general, descansa en la Teoría de las Emociones Básicas de Paul Ekman (1992). Esta teoría asume que existen emociones básicas (ira, miedo, alegría, tristeza, y asco) y universales que pueden identificarse a través de la expresión facial (para una revisión crítica más extensa ver Durán et al., 2017; Durán y Fernández-Dols, 2021). Si bien estas emociones se reflejan en el rostro, 'delatando' la emoción real, una persona podrá hacer intencionalmente que sus expresiones visibles sean discordantes con lo que realmente siente (Crivelli y Fridlund, 2019). En el marco de esta teoría se deriva una hipótesis para la evaluación de testigos (no del testimonio): si la expresión facial no concuerda con su estado emocional, la persona está mintiendo sobre dicho estado emocional (Patterson et al., 2023). La expresión facial completa revela la emoción que la persona quiere mostrar, pero la auténtica emoción se filtra en forma de micro expresión (hipótesis del filtraje, Ekman y Friesen, 1969; Vrij et al., 2019), una expresión emocional involuntaria y fugaz que dura entre 0.04 y 0.20 segundos (Matsumoto y Hwang, 2018).

La práctica de esta teoría en el contexto forense resulta altamente desaconsejable por los mismos motivos que ya se han abordado en anteriores apartados, pero, además, porque el concepto de 'congruencia o coherencia emocional' es una premisa particularmente insidiosa. Dicha premisa parte de la existencia de una 'línea base emocional' esperable en cada situación determinada, o para cada tipología de víctima o delito. Si por ejemplo se administra el citado FEAP para evaluar la coherencia emocional de una víctima de agresión sexual o violencia de género, se presupone que se dispondrá de evidencia del perfil emocional de esta tipología de víctima para poder establecer qué es esperable y qué no. No obstante, este planteamiento no se ve respaldado por la bibliografía científica. Hoy en día no se dispone de evidencia, ni nacional ni internacional, del perfil emocional de ningún tipo de víctima, en ningún tipo de contexto, y no es posible establecer qué es esperable, más allá de hacer inferencias derivadas de una opinión subjetiva. Las emociones esperables planteadas en este tipo de prácticas están basadas en razonamientos estereotipados sobre cómo debe comportarse el testigo-víctima que, además de no tener respaldo científico, pueden ser totalmente erróneos e idiosincrásicos (es decir, con interpretaciones contradictorias). Por ejemplo, en este tipo de informes puede analizarse la honestidad de la sonrisa, y podrían leerse afirmaciones altamente cuestionables como 'la presunta víctima genera sonrisas incongruentes con un abuso sexual'. Para el contexto forense, esa afirmación requiere solventar una serie de preguntas para las cuales la bibliografía científica no tiene respuesta: ¿Cuántas sonrisas incongruentes deben producirse para establecerse la incoherencia emocional? ¿Cómo se combina el número de sonrisas incongruentes con el resto de las emociones incongruentes para establecer una estimación de incoherencia emocional? ¿La presentación de la sonrisa difiere en función de variables personales, del diagnóstico, del evento traumático, etc.? ¿Qué tasa de error hay asociada a estos indicadores de manera individual y general? Además, como ocurre con las micro expresiones, el estudio de la sonrisa presenta una base teórica cuestionable. De facto, se sustenta en la hipótesis de la sonrisa de Duchenne, según la cual las sonrisas que son producto de una emoción genuinamente positiva incluyen la constricción de los ojos, y permite diferenciar cuando una sonrisa es genuina o falsa (Krumhuber y Manstead, 2009). A pesar de gozar de cierta aceptación, el apoyo empírico de esta hipótesis ha sido mixto, se ha observado que pueden ocurrir cuando una persona experimenta emociones negativas (Harris y Alvarado, 2005; Papa y Bonanno, 2008), que es posible fingirlas (Krumhuber y Manstead, 2009; Schmidt et al., 2009) y que, en general, más que un indicador de una emoción positiva es más compatible con un artefacto de una sonrisa intensa (Girard et al., 2021). No obstante, aun en el supuesto de que la sonrisa de Duchenne permitiera distinguir sonrisas genuinas de falsas, se debería poder responder a otra pregunta vital: ¿Qué implica una sonrisa genuina o falsa en, por ejemplo, una presunta víctima de violencia de género? Así pues, no es posible establecer este indicador, y ningún otro, como evidencia de 'incongruencia/incoherencia emocional'. Cualquier inferencia realizada en este sentido carecerá totalmente de respaldo cien-

tífico y, por ello, será inadmisible en un procedimiento judicial.

Múltiples indicadores e instrumentos: Técnicas complementarias y protocolos

Otra idea cuestionable es que la práctica del CNV suele recomendarse como una herramienta complementaria a otras técnicas o instrumentos y se presenta como un 'peritaje transversal', especialmente para los casos de credibilidad del testimonio. Se plantea que el análisis del CNV no es el elemento principal para la toma de decisiones, sino una técnica que aporta información adicional en una estrategia multi-método. Además, los expertos en CNV instan a no fijarse exclusivamente en aquellos indicadores que indiquen 'mentira', si no en todos los posibles que puedan obtenerse a lo largo del testimonio. Este planteamiento suele apoyarse en lo expresado habitualmente en los manuales de psicología forense (Carrasco-Ortiz y Rubio-Garay, 2020; Dujo-López et al., 2022; Sierra et al., 2010), donde se resalta la importancia de desarrollar una valoración psicológica basada en múltiples fuentes de información. De este modo, resulta frecuente encontrar protocolos, 'meta-protocolos' o sistemas de evaluación que proponen combinar múltiples instrumentos o herramientas, generalmente técnicas de análisis de contenido y técnicas de CNV (Vrij, 2015). Si bien la propuesta puede parecer apropiada, lo cierto es que plantea más inconvenientes que ventajas.

Cuando dos o más técnicas, instrumentos o herramientas se combinan para evaluar una misma variable, ya sea honestidad, credibilidad, invalidez de síntomas, riesgo de reincidencia, coherencia emocional, etc., debe disponerse de reglas de decisión para establecer el estado de la variable en cuestión (Larrabe et al., 2019; Vrieze y Grove, 2010) y ha de incrementar la validez de la medida (Arce, 2017). Por muy organizado y sistemático que sea un protocolo, o 'meta-protocolo', la formulación de hipótesis o conclusiones que no estén respaldadas por criterios verificados estadísticamente no tienen ningún tipo de utilidad para la práctica forense, dado que no hay posibilidad de determinar su grado de precisión, y por ello, de error. Planteemos, por ejemplo, una presunta valoración de la credibilidad en la que se han utilizado tres componentes: análisis de contenido de la declaración, análisis facial y análisis corporal ¿Qué regla de decisión se sigue para determinar la credibilidad? ¿Cómo se valoran los resultados de dicha combinación? ¿Qué debe hacer el profesional si el análisis del contenido indica que hay suficientes criterios de credibilidad, pero el análisis fácil y corporal resulta en incongruencia emocional? ¿Qué precisión y tasa de error hay asociada a cada decisión? Asimismo, la suma de técnicas de CNV a otras no incrementa validez al no tener respaldo científico de eficacia en la detección de la mentira (y menos de la honestidad del testigo). Muy por el contrario, incrementa el ruido (tasa de error).

Algo similar ocurre con el uso de múltiples indicadores de CNV. En la bibliografía se reitera que la toma de decisiones debe ir derivada de la combinación de múltiples indicadores no verbales (Hartwig et al., 2014; Matsumoto y Wilson, 2023). Si bien recientemente se han desarrollado estudios experimentales prometedores sobre qué indicadores de CNV utilizar en conjunto, y que precisión estimada deriva de dicho uso conjunto (Matsumoto y Hwang, 2018, 2020), hoy en día no se dispone de evidencia científica robusta sobre cómo combinarlos de manera eficaz, y en qué regla de decisión basarse para alcanzar una determinada conclusión o hipótesis. Así pues, en protocolo que combine, por ejemplo, los múltiples canales expresivos del CNV (expresión facial, gestos, postura, orientación y movimiento, paralenguaje, proxémica, etc.) ¿Qué regla de decisión deberá tomar para dicha combinación? ¿Cuántos datos serán necesarios para formular una determinada hipótesis? ¿Qué ocurre si se obtiene evidencia discordante? Y más importante ¿Qué proporción de indicadores debe considerarse para estimar que hay 'señal de alarma'? No se debería alcanzar la misma conclusión con un 10 % de indicadores de supuesta incoherencia que con un 50 %, o un 80 %.

Otra cuestión de gran importancia en la combinación de múltiples técnicas o instrumentos es que la evidencia a favor del uso de los elementos individuales no debe tomarse como evidencia a favor del uso del conjunto. O, dicho de otra manera, que los componentes tengan 'validez científica' por sí mismos no implica que la tenga la técnica. La lógica puede indicar que el uso de múltiples instrumentos para evaluar una misma variable será un enfoque más robusto y seguro, pero lo cierto es que este pro-

ceder incrementa el riesgo de falsos positivos (en la clasificación de un testigo como honesto, o falso negativo en la clasificación de un testigo como mentiroso) debido a que la ratio de error se acumula con la administración de cada instrumento adicional (Bender y Frederick, 2018; Larrabe et al., 2019). Debe analizarse estadísticamente qué condiciones debe cumplir el protocolo para poder medir la variable de interés de manera satisfactoria. Por ejemplo, en el área de simulación de síntomas esta cuestión ha sido tratada en profundidad con los test de validez de síntomas y rendimiento (Erdodi, 2022; Larrabee, 2014; Larrabee, 2022; Larrabee et al., 2019). Si bien aún se trata de un tema en desarrollo, se ha alcanzado cierto consenso en que 2 resultados positivos en dos test diferentes de validez de síntomas o rendimiento (con un mínimo de 90 % de especificidad) en un protocolo de hasta 14 instrumentos, permiten identificar una presentación inválida o bajo rendimiento con un grado de certeza aceptable (Sweet et al., 2021).

Así pues, la propuesta de usar técnicas como elementos complementarios, o combinar técnicas o instrumentos, sin conocer su rendimiento individual y general, plantea un severo problema operativo. En el caso del CNV, no se dispone de reglas de decisión ni para combinar los elementos que componen la técnica (por ejemplo, cómo combinar indicadores no verbales), ni para combinarla en conjunto con otros instrumentos o técnicas (por ejemplo, con las de análisis de contenido). Por ello, no es posible mencionarse ni sobre su capacidad de clasificación ni sobre su tasa de error, lo que imposibilita su uso como técnica complementaria en el proceso judicial. Si bien cabe la posibilidad de seguir un sistema de decisiones basado en el juicio clínico (el juicio clínico no es científico), éste puede presentar sesgos y resulta imposible ofrecer una justificación basada en datos cuantitativos de la decisión tomada (Dror y Scurich, 2020; Faust y Aherm, 2012; Garb, 2005). Además, aun en el supuesto de que no se considere como binario el estado de la variable (e.g., creíble vs. no creíble) y se plantee como un apoyo en un proceso de contraste de hipótesis, debe conocerse la tasa de error asociada a la confirmación/rechazo de cada una de las hipótesis. Que la técnica o herramienta se plantee como un elemento periférico cuyo resultado no tiene que valorarse en solitario no debería utilizarse para justificar la introducción en la valoración forense de prácticas cuyo impacto, positivo o negativo, no es posible establecer con seguridad.

Conclusiones y recomendaciones para la práctica profesional

A través de la presente revisión del estado del arte se puede concluir que la evidencia sobre el uso del CNV en el contexto jurídico-forense es tremendamente limitada, especialmente en población española, y no cumple con los estándares *Daubert* de admisibilidad de la prueba científica, es decir, judicialmente es una prueba pseudocientífica. Científicamente, la práctica totalidad de los indicadores de CNV estudiados (DePaulo et al., 2003; Sporer y Schwandt, 2006, 2007) presentan un tamaño del efecto no significativo. Para revertir estos resultados a un efecto significativo, sería necesario un número muy elevado de estudios con efectos significativos en cada uno de los indicadores de CNV. Y, dado que no es posible, la conclusión sobre la falta de validez de los indicadores CNV es definitiva. Además, los modelos predictivos son contradictorios: un modelo puede predecir un resultado para un indicador y otro justamente lo contrario (error de idiosincrasia). Esta contradicción en la predicción también anula (falta de persistencia en el contexto judicial, y de consistencia en el científico) de forma inequívoca y definitiva la validez de la medida.

En suma, el uso del CNV en la práctica pericial forense carece de validez judicial y científica. Esto se aplica a todas las prácticas pseudocientíficas en las que se ha refugiado la disciplina tanto en la detección de la mentira o el engaño, como en procesos de credibilidad del testimonio en cualquier tipología de delito, su uso para determinar la coherencia o congruencia emocional, o su uso en general como técnica complementaria o transversal para el apoyo de otros informes periciales.

La aplicación del CNV en el contexto jurídico-forense, y de seguridad en general, es una disciplina joven y en vías de desarrollo a la que se le ha otorgado una aplicabilidad profesional que no ha alcanzado en ningún momento, ni alcanzará en el futuro. Además, se ha recomendado el uso de técnicas que estaban en un estado prematuro y experimental, descuidando en el proceso el establecimiento de

un marco teórico robusto y una práctica científica sólida en la cual apoyarse. Dadas las severas consecuencias que tienen las valoraciones periciales en el contexto jurídico-forense, seleccionar los instrumentos de evaluación apropiados es una responsabilidad fundamental de los profesionales del ámbito (DeMatteo et al., 2019). Toda conclusión alcanzada en una valoración pericial debe estar correctamente respaldada por el sistema de evaluación administrado, y fundamentada por evidencia cuantitativa que permita valorar el grado de precisión de las estimaciones, así como su tasa de error.