Revista Iberoamericana de Psicología y Salud (2025) 16(2) 95-107

Revista Iberoamericana de Psicología y Salud

Iberomerican Journal of Psychology and Health

https://www.rips.cop.es/ • ISSN 2171-2069

Revista Oficial de la Federación Iberoamericana de Asociaciones de Psicología (FIAP) - Official Journal of the Latin-American Federation of Psychological Associations

Article

Estimating the Efficacy of the Gough's F–K Index to Detect Simulation: A Meta-Analytic Review.

Martín Pouso[1], Lorena Miranda[1], Noelia Souto[1] and Dolores Seijo[1,2]

[1] Unidad de Psicología Forense, Facultad de Psicología, Universidad de Santiago de Compostela (Spain).
[2] Departamento de Ciencia Política y Sociología, Facultad de Psicología, Universidad de Santiago de Compostela (Spain).

ARTICLE INFO

Received: 12/02/2025

Accepted: 02/06/2025

Keywords:
Forensic assessment
Differential diagnosis of simulation
Validity scales
Deception
Experimental design

ABSTRACT

Background/Objectives: The Gough's F–K index is one of the most effective validity indicators of the MMPI-2 (a reference instrument in the forensic setting) for discriminating between simulated and genuine profiles. However, the reported effect sizes are so large that they may be attributable, to some extent, to experimental design artifacts. Therefore, this study aimed to estimate the true effect size of the F–K Index in discriminating between simulated and genuine profiles, as well as the effect of the comparison group type (in-study control group vs. combined control group). **Method:** A total of 54 studies were identified, from which 77 effect sizes were obtained. A psychometric random-effects meta-analysis was performed on experimental effect sizes. **Results:** The results showed an average true effect size of $\delta = 3.67$, which was positive, very large, significant, and generalizable for the in-study control group comparison condition. Likewise, the results showed an average true effect size of $\delta = 2.43$, which was positive, very large, significant, and generalizable for the combined control group comparison condition. Comparatively, the results indicated a significantly larger effect for the in-study control group comparison condition. **Discussion:** The true effect found using data from the original studies (i.e., comparison with the in-study control group) is so large (> 99th percentile of all possible values) that it is undoubtedly overestimated. A source of this overestimation is the use of a non-normative control group configuration. The implications of these findings for the forensic evaluation of simulation in the forensic context are discussed.

Estimando la Eficacia del Índice F–K de Gough en la Detección de la Simulación: Una Revisión Meta-Analítica

Palabras clave:
Evaluación forense
Diagnóstico diferencial de simulación
Escalas de validez
Engaño
Diseño experimental

RESUMEN

Antecedentes/Objetivos: El índice F–K de Gough es uno de los indicadores de validez del MMPI-2 (instrumento de referencia en el ámbito forense) con mayor efectividad en la discriminación entre protocolos de respuesta simulados y genuinos. Sin embargo, los resultados son tan elevados que pueden responder en cierta medida a efectos del diseño experimental. Por ello, nos planteamos un estudio con el objetivo de estimar el tamaño del efecto verdadero del Índice F–K en la discriminación entre protocolos de respuesta simulados y genuinos, así como el efecto debido al grupo de comparación (grupo control del propio estudio vs. grupo control combinado). **Método:** Se encontraron un total 54 estudios, de los que se obtuvieron 77 tamaños del efecto. Con los tamaños de efecto experimentales se ejecutó un metaanálisis psicométrico de efectos aleatorios. **Resultados:** Los resultados mostraron un tamaño del efecto verdadero promedio, $\delta = 3.67$, positivo, más que grande, significativo, y generalizable en la condición de comparación con el grupo control del propio estudio. Asimismo, los resultados mostraron un tamaño del efecto verdadero promedio, $\delta = 2.43$, positivo, más que grande, significativo, y generalizable en la condición de comparación con el grupo combinado. Comparativamente, los resultados informaron de un efecto significativamente mayor en la condición de comparación con el grupo control del propio estudio. **Discusión:** El efecto verdadero con los datos de los estudios originales (comparación con el grupo control del propio estudio) es tan grande (> que el 99% de todos los posibles) que es evidente que esta sobrestimado. Una fuente de esta sobrestimación es la configuración del grupo control que no es normativo. Se discuten las implicaciones de estos resultados para evaluación forense de la simulación en el contexto forense.

Cite as: Pouso, M., Miranda, L., Souto, N., & Seijo, D. (2025). Estimating the efficacy of the Gough's F–K index to detect simulation: A meta-analytic review. *Revista Iberoamericana de Psicología y Salud, 16*(2), 95-107. https://doi.org/10.70478/rips.2025.16.10

Correspondence author: Dolores Seijo. E-mail: mariadolores.seijo@usc.es

Introduction

The Declaration of Basic Principles of Justice for Victims of Crime and Abuse of Power (United Nations [UN], 1985) states that "'victims' means persons who, individually or collectively, have suffered harm, including physical or mental injury, emotional suffering, economic loss or substantial impairment of their fundamental rights, through acts or omissions that are in violation of criminal laws operative within Member States, including those laws proscribing criminal abuse of power" (p. 1).

Thus, victim status necessarily requires the suffering of harm as a consequence of a crime defined in criminal law. However, this definition distinguishes several types of harm resulting from the victimization process, including mental injury or emotional suffering, that is, the psychological harm or imprint of the crime. In the forensic context, although crime victimization is compatible with numerous disorders, the literature has identified Post-Traumatic Stress Disorder (PTSD) as the primary imprint of psychological harm (Kessler et al., 1995, 2005), as it is considered the only one capable of establishing causality, rather than mere comorbidity, with the event under investigation. Adjustment Disorder (AD) is also contemplated when all criteria for PTSD are met, but the stressor (Criterion A in the DSM-5-TR; American Psychiatric Association [APA], 2022) is of a psychosocial nature; likewise, Acute Stress Disorder is considered when the duration of symptoms is less than one month (Arce, 2018).

The assessment of psychological harm, along with the assessment of witness memory (i.e., credibility) and/or psychological characteristics related to the capacity to testify, require a multimeasure (two measures) and multimethod (interview and psychometric measure) approach that includes control for (dis)simulation (Arce et al., 2006, 2009; Graham, 2011; Rogers, 2018b; Vilariño et al., 2013). In other words, in the criminal context, forensic psychological assessment must consist of an interview, psychometric instrumentation, and, when necessary, other additional tests (Arce & Fariña, 2012). Moreover, deception must be suspected to preserve the principles of legality and presumption of innocence, since an innocent person can never be wrongfully convicted (i.e., false positives must be 0; Sentence of the Spanish Supreme Court 1029/1997, December 29), whereas it is sufficient that guilty parties are generally convicted (i.e., false negatives are admissible, but reprehensible and unethical; Sentence of the Spanish Supreme Court 213/2002, February 14). In criminal forensic evaluations, this deception translates into the differential diagnosis of simulation (malingering), that is, ruling out the "intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives" (APA, 2022, p. 835), one of which is the evasion of criminal responsibility. Thus, if simulation cannot be ruled out, the psychological assessment is not valid as inculpatory forensic evidence (Arce, 2018). Furthermore, the simulation of harm in the forensic context is a common phenomenon in Spain (Puente-López et al., 2023, 2024), and its differential diagnosis is compatible with other alternative hypotheses, such as the severity of the harm (Arce, 2017; Arce, Fariña, & Vilarino, 2015; Graham, 2011; Fariña et al., 2014). Therefore, it must be ruled out before assessing the psychological harm.

Based on this premise, not just any method is acceptable, the techniques used in a forensic assessment must meet the Daubert criteria of scientific validity (Daubert v. Merrell Dow Pharmaceuticals, 1993). In this regard, the clinical-forensic interview (Arce & Fariña, 2001; Vilariño et al., 2013) is posited as an appropriate method for establishing a causal relationship between the psychological harm and the event under investigation. However, it does not discriminate against all forms of simulation; that is, it requires a complementary detection task: psychometric instrumentation (Rogers, 2018a). Several valid tests exist for the forensic context, such as the SCL-90-R (Vilariño et al., 2020) or the MMPI-2 (Gancedo et al., 2021; Graham, 2011; Lees-Haley et al., 2002; Rogers, 2018a; Rogers et al., 2003). The latter is the benchmark instrument and the most widely used in forensic assessment, as it measures psychopathology and psychological adjustment, and features validity scales and indices that rule out, in this case, all forms of simulation (Pope et al., 2006; Puente-López et al., 2024).

The Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher et al., 1989; Butcher et al., 2019) assesses mental health through 9 clinical scales (there were originally 10, but the fifth scale, Masculinity-Femininity [Mf], is now obsolete due to the absence of mental disorders that include masculinity or femininity in their symptomatology): Scale 1 (Hs: Hypochondriasis), Scale 2 (D: Depression), Scale 3 (Hy: Conversion Hysteria), Scale 4 (Pd: Psychopathic Deviate), Scale 6 (Pa: Paranoia), Scale 7 (Pt: Psychasthenia), Scale 8 (Sc: Schizophrenia), Scale 9 (Ma: Hypomania), and finally, Scale 0 (Si: Social Introversion). On the other hand, for controlling simulation, the MMPI-2 features several validity scales. In the forensic context, these include not only the original scales from the commercial version (i.e., ?, F, K, L) but also additional ones (i.e., TRIN, VRIN, Fb, Fp, FBS, RBS), as well as other widely validated scales and indices (F–K [Gough Index], F-Fb, %True responses, Fs, Ds, Fptsd) that demonstrate good discriminant validity and high classification accuracy for simulation (Gancedo, Sanmarco et al., 2020; Gancedo, Selaya et al., 2020; Gancedo et al., 2021; Rogers et al., 2003).

In particular, the Gough Index (F–K) is established as a validity indicator for the MMPI-2 (Brophy, 1995; Lees-Haley, 1989, 1991; Osborne et al., 1986; Sánchez et al., 2008), although it is not included in the commercial version following the restandardization by Butcher et al. (1989), nor in the Spanish adaptation (Butcher et al., 2019). It is composed of a total of 90 items on the MMPI-2 (it does not appear in the MMPI-2-RF or MMPI-A): 60 items from the F scale, which aims to detect infrequency or incoherence in responses (Gancedo et al., 2021; Greene, 2011; Rogers, 2018a); and 30 items from the K scale, whose function is to detect a subtle but persistent tendency to either exaggerate psychopathology and present in a very unfavorable light (low scores) or to deny psychopathology and present in a favorable light (high scores) (Graham, 2011; Fariña et al., 2017). This allows the Gough Index to interpret results from a dual dimension, such that if the raw score from the F–K difference is negative (F–K < -15/-21), dissimulation is suspected (Greene, 2011; Fariña et al., 2017), whereas if the F–K difference is positive (F–K ≥ 30), simulation is suspected (Arce et al., 2009; Rogers 2018a).

Originally, Gough (1947, 1950) proposed the F–K index as a measure of feigned emotional distress, and although many cut-off scores have been proposed, Graham (1990) suggests that as the F–K score increases above zero, the probability of simulation also increases. In fact, Gough (1950) found that an F–K score > 2 correctly detected 86% of faked MMPI profiles while only misidentifying 11% of genuine profiles. He therefore proposed the values F–K < -9 for the tendency to present a "good image" (i.e., dissimulation) and F–K > 9 for the tendency to feign a "bad image" (i.e., simulation), considering a profile valid only if its raw score fell within that range. However, these cut-off scores have been the subject of constant debate in the scientific community and have thus been corrected over the years (Fox et al., 1995). Nevertheless, no single index or scale classifies deception on its own; only their combination increases the validity of the approach for controlling false positives (Arce et al., 2024). Therefore, to validate a protocol and ensure a high probability of response distortion, a high Gough Index (F–K ≥ 30) is required, along with two or more of the following criteria: F, Fb, or Fp ≥ T-score of 66.45 (Arce et al., 2009; Rogers, 2018a). In this context, only a positive F–K score is of interest—that is, the measure of simulation—as dissimulation has little to no relevance in the criminal field.

With the foregoing in mind, it can be concluded that the MMPI-2 is a widely studied and generally validated tool in the forensic context as a complementary test (Arce et al., 2009; Pope et al., 2006). However, elements remain to be explored. While numerous studies have aimed to detect the simulation of harm using the MMPI-2 validity scales and indices, there are scarcely any meta-analytic reviews oriented toward verifying their efficacy, especially regarding the F–K index (Nelson et al., 2006, 2010; Rogers et al., 1994). Moreover, until recently, quantitative syntheses did not consider the study design type as a moderator variable, despite its influence on the obtained results and, therefore, on their generalization to forensic practice (Gancedo, Novo, & Arce, 2020). They only considered the comparison sample and the specific disorder being simulated (Rogers et al., 2003; Sharf et al., 2017). In this regard, it has been shown that many empirical studies in this field do not use a normative population (understood as the general population) for comparison; instead, they often use in-study groups in which an attempt is made to match the sociodemographic characteristics of the groups in question (i.e., experimental and control; Gancedo, Novo, & Arce, 2020).

For all these reasons, a meta-analytic study was designed with the objective of estimating the true effect size from experimental studies that tested the discriminant efficacy of the Gough Index (F–K) in distinguishing between simulated and genuine response protocols, as well as to analyze the effects of the control group type (in-study vs. normative). Consequently, we propose to test the following hypotheses:

Hypothesis 1: The F–K Index on the MMPI-2 significantly discriminates between genuine and simulated response protocols.

Hypothesis 2: The effect of comparing simulation responses with the in-study control group will be significantly larger than the effect of comparing them with a combined (normative) control group.

Hypothesis 3: The validity of the results (discriminant capacity between genuine and simulated response protocols) is generalizable across studies.

Method

Literature Search

To obtain the largest possible number of studies that quantified the ability of the F–K Index on the MMPI-2 to discriminate between genuine and simulated responses, or that published data allowing for the calculation of this effect, several strategies were employed. First, a search was conducted in the reference scientific electronic databases: Scopus, Web of Science, and PsycInfo. The following descriptors were used as meta-tags: *MMPI AND* (Gough index OR F–K index) FOR the F minus K variable; AND ("fake bad" OR "malinger" OR "faking bad" OR "simulation" OR "overreporting") FOR the simulation variable*. Second, the reference sections of the articles found in the database were reviewed. Third, references contained in previous systematic and meta-analytic reviews that included the Gough Index on the MMPI-2 as a measure of simulation were identified: Nelson et al. (2006, 2010) and Rogers et al. (1994). Although an exact replication of existing reviews would be ideal, this task was not always possible due, on the one hand, to access limitations for the retrieved works and, on the other hand, to the non-inclusion of the necessary primary data in said studies. Finally, expert researchers in the field were contacted to obtain unpublished data.

Once the studies were identified and collected, they were reviewed. The first step consisted of detecting duplicate records (i.e., database entries, not studies), for which automated tools were used, although a manual procedure was also performed. Next, an initial screening by title and abstract was performed, and subsequently, the resulting studies were assessed according to the following inclusion criteria: the instrument had to be the MMPI-2; it had to report sociodemographic characteristics of participants relevant for the review and the study of moderators (number of participants, age, gender [as norms vary between gender], etc.); it had to provide an effect size calculation or statistical data that allowed for its calculation, such as means and standard deviations or the classification rate for simulation.

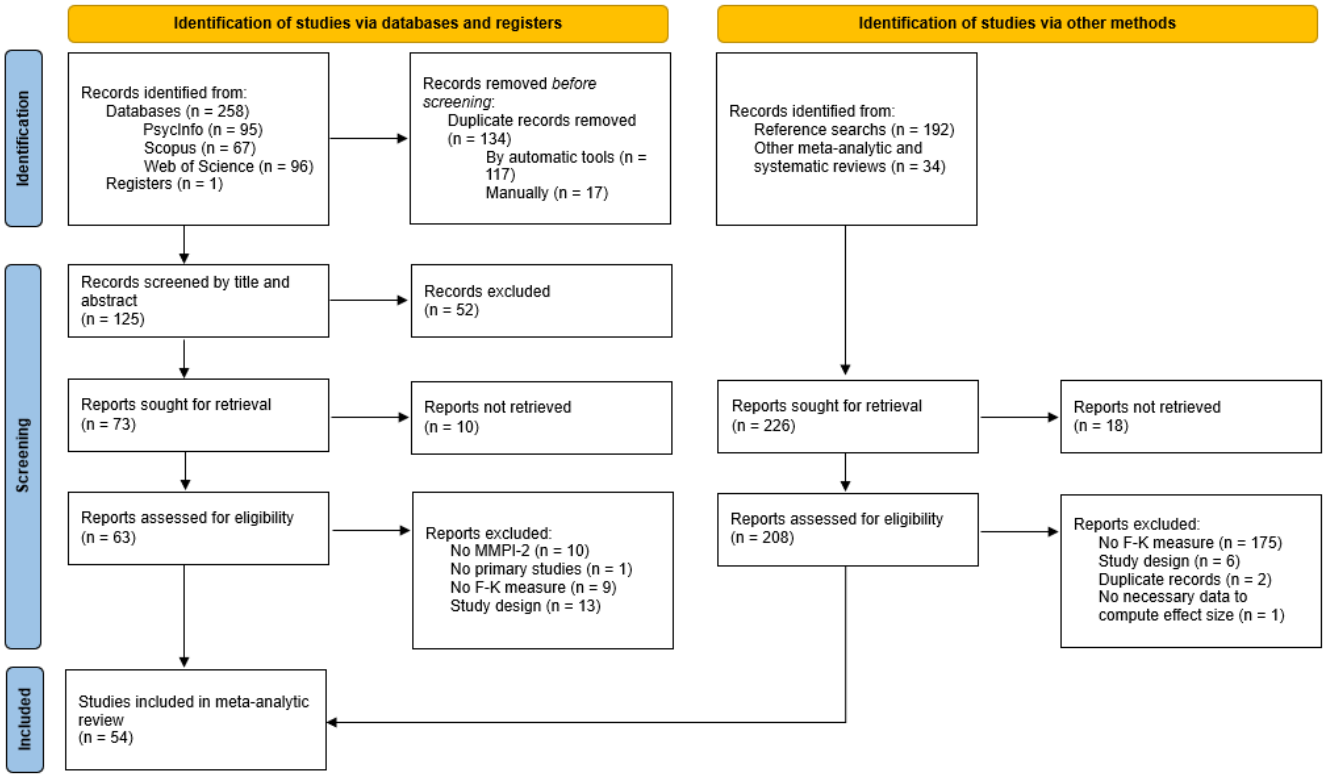
a. Study design (Rogers, 2018c): simulation research (SR); known-groups comparison (KGC); differential prevalence design (DPD); and bootstrapping comparison (BTS).
b. Design type: group comparison with a simulation group.

The following exclusion criteria were applied: a) studies based on the same dataset (resulting in duplicate effects), a common practice in this field, were excluded as they are not statistically independent; and b) unpublished studies or those published without peer review, or, in the case of doctoral dissertations, not defended before a panel (Daubert standard), were excluded.

Figure 1 presents the flowchart for the study search process for the present quantitative synthesis, developed following the PRISMA methodology (Page et al., 2021). This procedure yielded a total of 54 documents (see Annex), from which 77 effect sizes were obtained for the F–K Index: 53 of them from the comparison between Simulation Groups and the combined Control Group, and 24 related to the comparison between simulators and their respective in-study Control Group.

Figure 1

Flow diagram of the meta-analysis



Coding of the Primary Studies

Once the primary studies included in the meta-analytic review were selected, the information considered relevant from each study was coded. The following categories were coded: a) the primary study reference; b) document type (peer-reviewed paper, published doctoral dissertation); c) design characteristics (i.e., the moderator variables introduced); d) sample characteristics (i.e., size, sex, age, group of origin); e) the mean and standard deviation of the groups to be compared or, failing that, the necessary data to calculate the effect size; and f) the effect size (the reported effect sizes were recalculated with the data from the article itself to verify their accuracy).

The coding procedure was carried out by two independent, trained, and experienced raters (Montes et al., 2022). Between- and within-rate concordance was estimated using true Kappa ($\bar{k}$; Fariña et al., 2002); that is, the concordance was corrected by the correspondence between raters, thus avoiding the coding of two errors as an agreement (Arce et al., 2000). For all studies, the results showed a total true between-rater concordance ($\bar{k} = 1$). Both raters recoded the protocols one week after the original coding (within-rater concordance), obtaining a total true concordance ($\bar{k} = 1$). In short, the raters demonstrated stability in their coding over time (test-retest) and between each other. Consequently, other equally trained raters would have coded the studies in the same way on the analyzed variables, meaning they coded the variables with accuracy according to the coding categories (i.e., coding fidelity; Fariña et al., 2002).

Data Analysis

All effect sizes were calculated based on the *d* estimator from the data provided by the primary studies (i.e., using the sample sizes and the means and standard deviations of the corresponding group). This was justified because a large percentage of the studies did not report an effect size, or, if they did, they did not report how it had been computed. Therefore, the formula of Cohen (1988; for when groups are of equal size and for comparison with a test value), Hedges (1981; for when groups differ in size), or Glass (1977; for when the standard deviations vary considerably, i.e., the assumption of homogeneity of variances is violated, as occurs in the computation of the combined control group) was applied. With this information, Excel spreadsheets were created to ensure precise calculations, and their operational accuracy was verified by comparing them with manual execution. Next, the existence of outliers was examined by checking for the presence of *outliers* (±1.5*IQR) and extreme values (±3*IQR), as well as compliance with Chauvenet›s criterion (1891; ±2SD).

On the other hand, to quantitatively aggregate the results of the primary studies, the psychometric meta-analysis procedure for standardized differences (i.e., experimental effect sizes) by Schmidt and Hunter (2015) was followed, correcting the effect size for sampling error and for the unreliability of the criterion (dependent variable).

The significance of the meta-analytic results was estimated with Z, and the comparisons between the meta-analytic results were made with the *q* statistic, with its significance also tested with Z (Arce et

al., 2023; Cohen, 1988). Finally, the effect sizes were interpreted qualitatively in terms of small (d = 0.20), moderate (d = 0.50), large (d = 0.80), and very large (d > 1.20) magnitude (Arce, Fariña, Seijo et al., 2015; Cohen, 1988), and quantitatively with the Probability of Superiority of the Effect Size (PS_{ES}; Arce et al., 2020; Arias et al., 2020).

Criterion Reliability

Since the Gough Index is a combination of two scales (i.e., F and K), neither the primary studies nor the original manuals provide data on its reliability. Therefore, for its calculation, the internal consistency (i.e., Cronbach's α; 1951) of both scales was taken as the unit of measurement, and an empirical distribution of reliability coefficients was created for each of the analyzed variables based on the obtained data. Thus, for both the F scale and the K scale, data from Butcher et al. (2019), García & Romero (2009), Mundia (2011), Silberman (2002), and Tarescavage et al. (2015) were used. Based on these values, the formula developed by Mosier (1943) for the reliability of a composite was applied, yielding an internal consistency (α) of .89.

Study of Moderator Variables

A moderator variable is one that "causes differences in the relationship between two other variables" (Hunter & Schmidt, 2004, p. 90). In other words, it is a particular feature of the studies that varies from one to another and can affect the results (in the form of non-artifactual variance). Consequently, Schmidt and Hunter (2015) recommend breaking down the set of studies and creating smaller subcategories based on the characteristic to be analyzed. In this study, the data were coded considering three moderator variables, which have been previously considered (Gancedo, Novo et al., 2020): (1) the comparison population, (2) the simulated harm, and (3) the study design.

1. *Comparison population.* This is a moderator studied in other reviews with a reported effect on the effect size (Gancedo, Novo et al., 2020; Gancedo, Selaya et al., 2020). While there are several relevant comparison groups, such as the clinical population (i.e., a positive criterion for non-simulation, or discriminant validity) and/or the general population (i.e., normative validity of the instrument or scale/index), the latter has greater transferability to forensic practice (Arce, 2017), since the goal of an assessment is to distinguish individuals who present with symptoms of psychological harm from the general population. However, it has been found that many studies using their own control group lack external validity (i.e., generalizability), as these groups are often analogous to the simulator group in sociodemographic characteristics and are not representative of the general population (Gancedo, Novo et al., 2020; Redondo et al., 2019). This leads to the observation of experimental effect sizes that are excessively larger than expected, that is, larger than the population mean intended for the test (see Annex).
2. *Simulated harm*. It is a fact that, depending on the context, it is interesting to know the ability of the scales to detect specific disorders (e.g., schizophrenia, PTSD, depression, brain damage; Rogers et al., 2003; Sharf et al., 2017). However, it is relatively common to find small samples in this type of study (N < 400), which diminishes the generalizability of the results (Gancedo, Selaya et al., 2020). Furthermore, the validity indices and scales of the MMPI-2 are mostly designed to assess simulation in general, not a specific disorder, unlike the MMDS (Henry et al., 2008).
3. *Study design*. Finally, different methodological approaches for measuring simulation have been found in the collected primary studies, which can affect both the results and their validity. Rogers (2018b) differentiates four types: 1) simulation research (SR), in which the experimenter randomly assigns participants to experimental conditions (e.g., to respond honestly or in a simulated manner); 2) known-groups comparison (KGC), in which participants are classified based on some external indicator (e.g., the result of a test); 3) differential prevalence design (DPD), in which participants are classified based on their prior membership in groups that are presumed to exhibit a particular response style, different from that of the normative population (e.g., litigants, clinical patients, etc.); and (4) bootstrapping comparison (BTS), in which indicators and detection strategies are used to distribute participants (e.g., variance, biases, etc.). In terms of efficacy, although the simulation design is the most common and is posited to have higher internal validity than its counterparts, it also lacks generalizability (especially in the forensic context). This is due to several reasons: first, the nature of the incentives and repercussions between an experimental group and a real-world group in a judicial context (Rogers et al., 2019); second, the different formulations of the instructions in the various studies (Merckelbach et al., 2009); and third, the different simulation strategies followed—that is, whereas in a forensic context the goal is to simulate successfully while avoiding detection, in an experimental context the success of the simulation is maximized without regard for the risk of detection (Fariña et al., 1994). Conversely, the design with the highest external validity (i.e., applicability in the forensic context) is the known-groups design, followed by the differential prevalence and bootstrapping designs.

For all the reasons stated above, the first and last of these moderator variables were taken into consideration for the purposes of the present study.

Results

Outlier Analysis

Prior to the meta-analytic study, it is necessary to perform an exploratory data analysis with the objective of identifying outliers or extreme values, which can distort the value of the observed mean effect size, defined as "observations which deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism" (Hawkins, 1980, p. 1). An exploratory data analysis of the observed effect sizes did not show outliers.

General Results of the Meta-Analyses

The results of the meta-analysis (k = 55) comparing the Simulation Group (N = 4931) with a combined Control Group (N = 3158) exhibited (see Table 1) a mean true effect size of δ = 2.43. This effect was positive (higher F–K scores in the simulation group), significant (Z = 8.64, p < .001), generalizable across studies (the credibility interval does not include 0), and of a very large magnitude (δ > 1.20; greater than 95.73% of all possible effects, PS_{ES} = .9573). These results confirm Hypotheses 1 and 3. Nevertheless, the distribution of the studies is heterogeneous (%Var < 75%), with this variability being explained by moderators.

Table 1

Psychometric meta-analysis of the F–K Index comparing the Simulation Group with the Combined Control Group ($N_{Control}$ = 3158).

K	N	d	SD_{Obs}	S^2_{SE}	SD_{Res}	δ	SD_{δ}	%Var	80% CI
53	4764	2.29	1.30	0.07	1.27	2.43	1.34	4.41	0.69,4.17

Note. *k*: Number of effect sizes; *N*: Total sample size; *d*: sample-size weighted observed effect size; SD_{Obs}: Sample size weighted observed standard deviation of *d*-values; S^2_{SE}: Variance due to sampling error; %Var = Percent of variance in observed d-values accounted for by artifacts; SD_{Res}: Standard deviation of d-values after removing sampling error variance; *δ*: Mean true effect size; SD_{δ}: Standard deviation of delta; 80% CI: 80% Credibility Interval.

The results of the meta-analysis of the experimental effect sizes comparing the Simulation Group with the in-study Control Groups showed (see Table 2) a mean true effect size of delta = 3.67. This effect was positive (higher F–K scores in the simulation group), significant (Z = 14.22, p = .001), generalizable across studies (the credibility interval does not include 0), and of a very large magnitude (delta > 1.20; greater than 99.53% of all possible effects, PS_{ES} = .9953). These results confirm Hypotheses 1 and 3. Nevertheless, the distribution of the studies is heterogeneous (%Var < 75%), with this variability being explained by moderators.

Table 2

Psychometric meta-analysis of the F–K Index comparing the Simulation Group with the in-study Control Group.

K	N	d	SD_{Obs}	S^2_{SE}	SD_{Res}	δ	SD_{δ}	%Var	80% CI
24	5566	3.46	1.43	0.04	1.41	3.67	1.50	2.12	1.74,5.16

Note. *k*: Number of effect sizes; *N*: Total sample size; *d*: sample-size weighted observed effect size; SD_{Obs}: Sample size weighted observed standard deviation of *d*-values; S^2_{SE}: Variance due to sampling error; %Var = Percent of variance in observed d-values accounted for by artifacts; SD_{Res}: Standard deviation of d-values after removing sampling error variance; *δ*: Mean true effect size; SD_{δ}: Standard deviation of delta; 80% CI: 80% Credibility Interval.

When comparing the results of the meta-analyses on the effect of simulation on the F–K index by control group type (in-study control group vs. combined control group), the results showed a significantly smaller effect size, $qs(N' = 5116) = -0.342$, $Z = 17.29$, $p < .001$, for the condition comparing with the combined control group (δ= 2.43 vs. δ = 3.67) than for the condition comparing with the in-study control group. This finding confirms Hypothesis 2.

Discussion

A meta-analytic review was conducted with the objective of estimating the true effect size of the Gough Index (F–K) and the effects of the control group type (in-study vs. normative) on this estimation. Limitations to its generalization emerge from the results of this review. First, the mean true effect size resulting from the comparison with the in-study control group is larger than 99.53% of all possible effects, meaning the overall accuracy (i.e., the probability that an individual will be correctly classified by the F–K Index) would be practically perfect. This result contradicts the unreliability of the F–K Index found in this review (α = .89). Consequently, this result indicates that a large part of the effect is not due to the construct being measured, but rather to the method. Second, the results of the meta-analysis comparing the simulation group with a combined control group (true normative sample) are also of such a high magnitude—greater than 95.73% of all possible effects—that they are impossible. Third, the results are for mean comparison, then it is not directly valid for forensic practice (N = 1). Fourth, the cut score for the classification of simulation is known, then the results may not be generalized to forensic field (target: classification of simulation).

From these results and limitations, the qualitative validity of the findings can be concluded: the F–K Index discriminates positively and significantly between simulated and genuine response protocols. However, they are not valid quantitatively: the effects are overestimated. One source of this method-based overestimation is the comparison group in the primary studies: these are not truly normative groups. In fact, the effect is significantly reduced when the effect size is calculated with the aggregated sample from the primary studies: the true normative sample of these studies. This implies that the comparison samples were not well-selected in the primary studies. Along these lines, Gancedo, Selaya et al. (2020) had already warned that many of these studies used university students for both the simulator and general population groups, thereby introducing a population bias. Nevertheless, the mean effect size corrected for this population bias remains so high that it is also untenable. Therefore, future literature should be oriented toward the search for other moderators (variance due to the method and not the measured construct; in this case, efficacy in the discrimination and classification of simulation) that have artificially inflated the effect. In this regard, Arce et al. (2023) have pointed out that a large part of the variance in the observed effect size is due to the design type of the primary studies, specifically simulation research (SR) designs.

Despite this, the results allow for practical implications to be drawn for forensic practice. First, it has been verified that the F–K Index of the MMPI-2 effectively discriminates between genuine and simulated response protocols, regardless of the comparison group used, although the power of its efficacy is exaggerated. This implies that it can be used as a reliable and valid indicator for the detection and classification of simulation. However, the margin of error is unknown, especially the Type II error (false negatives), which is not admissible in forensic assessment as it contravenes the principle of presumption of innocence—that is, the conviction of a falsely accused person (Sentence of the Spanish Supreme Court 1029/1997, December 29). Consequently, this indicator must be used in combination with other indicators and guided by a forensic technique that controls for Type II error.

References

References marked with an asterisk indicate studies included in the meta-analysis

American Psychiatric Association. (2022). *Diagnostic and Statistical Manual of Mental Disorders, Text Revision DSM-5-TR*. American Psychiatric Publishing. https://doi.org/10.1176/appi.books.9780890425787

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 the witness statements: Evaluation of the scientific and judicial validity of the hypothesis and the forensic proof]. *Acción Psicológica, 14*(2), 171-190. https://doi.org/10.5944/ap.14.1.21347

Arce, R. (2018). Evaluación del daño psicológico: Psicometría, entrevista y técnica forense [Assessment of psychological harm: Psychometry, interview and forensic technique]. In E. Carbonell, D. Pineda, & M. Novo (Eds.), *Psicología jurídica: Ciencia y profesión. Colección Psicología y Ley, Nº 15* (pp. 235-244). Sociedad Española de Psicología Jurídica y Forense. https://shorturl.at/jRISO

Arce, R., & Fariña, F. (2001). *Construcción y validación de un procedimiento basado en una tarea de conocimiento para la medida de la huella psíquica en víctimas de delitos: La entrevista forense* [Unpublished manuscipt]. Unidad de Psicología Forense, Universidad de Santiago de Compostela.

Arce, R., Arias, E., Novo, M., & Fariña, F. (2020). Are interventions with batterers effective? A meta-analytical review. *Psychosocial Intervention, 29*(3), 153-164. https://doi.org/10.5093/pi2020a11

Arce, R., & Fariña, F. (2012). La entrevista psicológico forense a niños, adultos y discapacitados. In S. Delgado, F. Bandrés, & A. Tejerina (Eds.), *Tratado de medicina legal y ciencias forenses: Vol. V. Pediatría legal y forense. Violencia. Víctimas* (pp. 795-817). Bosch.

Arce, R., Fariña, F., & Fraga, A. I. (2000). Género y formación de juicios en un caso de violación [Gender and juror judgment making in a case of rape]. *Psicothema, 12*(4), 623-628. https://www.redalyc.org/pdf/727/72712417.pdf

Arce, R., Fariña, F., Carballal, A., & Novo, M. (2006). Evaluación del daño moral en accidentes de tráfico: Desarrollo y validación de un protocolo para la detección de la simulación [Evaluating psychological injury in motor vehicle accidents (MVA): Development and validation of a protocol for detecting simulation]. *Psicothema, 18*(2), 278-283. http://www.psicothema.com/psicothema.asp?id=3210

Arce, R., Fariña, F., Carballal, A., & Novo, M. (2009). Creación y validación de un protocolo de evaluación forense de las secuelas psicológicas de la violencia de género [Creation and validation of a forensic protocol to assess psychological harm in battered women]. *Psicothema, 21*(2), 241-247. http://www.psicothema.com/pdf/3621.pdf

Arce, R., Fariña, F., Seijo, D., & Novo, M. (2015). Assessing impression management with the MMPI-2 in child custody litigation. *Assessment, 22*(6), 769-777. https://doi.org/10.1177/1073191114558111

Arce, R., Fariña, F., & Vilariño, M. (2015). Daño psicológico en casos de víctimas de violencia de género: Un estudio comparativo de las evaluaciones forenses [Psychological injury in intimate partner violence cases: A contrastive analysis of forensic measures]. *Revista Iberoamericana de Psicología y Salud, 6*(2), 72-80. https://doi.org/10.1016/j.rips.2015.04.002

Arce, R., Marcos, V., Sanmarco, J., & Fariña, F. (2024). Is the self-reported information by male sentenced of violence against women in the intervention phase valid? *Anuario de Psicología Jurídica, 34*, 23-30. https://doi.org/10.5093/apj2023a3

Arce, R., Selaya, A., Sanmarco, J., & Fariña, F. (2023). Implanting rich autobiographical false memories: Meta-analysis for forensic practice and judicial judgment making. *International Journal of Clinical and Health Psychology, 23*(3), 100386. https://doi.org/10.1016/j.ijchp.2023.100386

Arias, E., Arce, R., Vázquez, M. J., & Marcos, V. (2020). Treatment efficacy on the cognitive competence of convicted intimate partner violence offenders. *Anales de Psicología/Annals of Psychology*, *36*(3), 427-435. https://doi.org/10.6018/analesps.428771

*Austin, J. S. (1992). The detection of fake good and fake bad on the MMPI-2. *Educational and Psychological Measurement, 52*(3), 669-674. https://doi.org/10.1177/0013164492052003016

*Bagby, R. M., Nicholson, R. A., Buis, T., & Bacchiochi, J. R. (2000). Can the MMPI-2 validity scales detect depression feigned by experts? *Assessment, 7*(1), 55-62. https://doi.org/10.1177/107319110000700104

*Bagby, R. M., Rogers, R., & Buis, T. (1994). Detecting malingered and defensive responding on the mmpi-2 in a forensic inpatient sample. *Journal of Personality Assessment, 62*(2), 191-203. https://doi.org/10.1207/s15327752jpa6202_2

*Bagby, R. M., Rogers, R., Buis, T., Nicholson, R. A., Cameron, S. L., Rector, N. A., Schuller, D. R., & Seeman, M. V. (1997a). Detecting Feigned Depression and Schizophrenia on the MMPI-2. *Journal of Personality Assessment, 68*(3), 650-664. https://doi.org/10.1207/s15327752jpa6803_11

*Bagby, R. M., Rogers, R., Nicholson, R., Buis, T., Seeman, M. V., & Rector, N. (1997b). Does clinical training facilitate feigning schizophrenia on the MMPI-2? *Psychological Assessment, 9*(2), 106-112. https://doi.org/10.1037/1040-3590.9.2.106

*Berry, D. T. R., Adams, J. J., Clark, C. D., Thacker, S. R., Burger, T. L., Wetter, M. W., Baer, R. A., & Borden, J. W. (1996). Detection of a cry for help on the MMPI-2: An analog investigation. *Journal of Personality Assessment, 67*(1), 26-36. https://doi.org/10.1207/s15327752jpa6701_2

*Berry, D. T. R., Cimino, C. R., Chong, N. K., LaVelle, S. N., Ho, I. K., Morse, T. L., & Thacker, S. R. (2001). MMPI-2 fake-bad scales: An attempted cross-validation of proposed cutting scores for outpatients. *Journal of Personality Assessment, 76*(2), 296-314. https://doi.org/10.1207/S15327752JPA7602_11

*Berry, D. T. R., Wetter, M. W., Baer, R. A., Youngjohn, J. R., Gass, C. S., Lamb, D. G., Franzen, M. D., MacInnes, W. D., & Buchholz, D. (1995). Overreporting of closed-head injury symptoms on the MMPI-2. *Psychological Assessment, 7*(4), 517-523. https://doi.org/10.1037/1040-3590.7.4.517

*Bianchini, K. J., Etherton, J. L., Greve, K. W., Heinly, M. T., & Meyers, J. E. (2008). Classification accuracy of MMPI-2 validity scales in the detection of pain-related malingering: A known-groups study. *Assessment, 15*(4), 435-449. https://doi.org/10.1177/1073191108317341

*Blanchard, D. D., McGrath, R. E., Pogge, D. L., & Khadivi, A. (2003). A comparison of the PAI and MMPI-2 as predictors of faking bad in college students. *Journal of Personality Assessment*, *80*(2), 197-205. https://doi.org/10.1207/S15327752JPA8002_08

Brophy, A. L. (1995). Gough's F–K Dissimulation Index on the MMPI-2. *Psychological Reports, 76*(1), 158. https://doi.org/10.2466/pr0.1995.76.1.158

*Bury, A. S., & Bagby, R. M. (2002). The detection of feigned uncoached and coached posttraumatic stress disorder with the MMPI-2 in a sample

of workplace accident victims. *Psychological Assessment, 14*(4), 472–484. https://doi.org/10.1037/1040-3590.14.4.472

Butcher, J. N., Graham, J. R, Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2.* Pearson Assessments.

Butcher, J. N., Graham, J. R, Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2019). *MMPI-2. Inventario Multifásico de Personalidad de Minnesota-2*. TEA Ediciones.

*Cavenagh, N. A. (2008). *An analysis of the differential power of the Fake Bad Scale of the MMPI-2* [Doctoral dissertation, University of Nevada]. UNLV Retrospective Theses & Dissertations. 2810. https://doi.org/10.25669/wyx8-8gl6

*Chang, Y.-T., Tam, W.-C. C., & Chiang, S.-K. (2017). Detection of feigning psychosis with multiscale personality inventories: A simulation design pilot study in Taiwan. *Sage Open, 7*(3). https://doi.org/10.1177/2158244017734023

*Charles, T. L. Jr. (1999). *Usefulness of the Minnesota Multiphasic Personality Inventory-2 in detection of deception in a personal injury type forensic population* [Tesis doctoral, Texas A&M University]. Commerce ProQuest Dissertations Publishing. https://www.proquest.com/psycinfo/docview/304573882/fulltextPDF/4E0CA1D8?sourcetype=Dissertations%20&%20Theses

Chauvenet, W. (1891). *A Manual of spherical and practical astronomy: 2: Theory and use of astronomical instruments* (Vol. 2, 5th ed.). Lippincott. https://shorturl.at/G6lx7

Cohen, J. (1988). *Statistical power analysis for behavioral sciences* (2nd ed.). LEA. https://doi.org/10.4324/9780203771587

*Cramer, K. M. (1995). The effects of description clarity and disorder type on MMPI-2 fake-bad validity indices. *Journal of Clinical Psychology, 51*(6), 831-840. https://doi.org/10.1002/1097-4679(199511)51:6<831::AID-JCLP2270510616>3.0.CO;2-O

*Crawford, E. F., Greene, R. L., Dupart, T. M., Bongar, B., & Childs, H. (2006). MMPI-2 assessment of malingered emotional distress related to a workplace injury: A mixed group validation. *Journal of Personality Assessment, 86*(2), 217-221. https://doi.org/10.1207/s15327752jpa8602_11

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

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993). https://supreme.justia.com/cases/federal/us/509/579/case.pdf

*Elhai, J. D., Gold, P. B., Frueh, B. C., & Gold, S. N. (2001a). «Cross-validation of the MMPI-2 in detecting malingered posttraumatic stress disorder»: Errata. *Journal of Personality Assessment, 77*(1), 189. https://doi.org/10.1207/S15327752JPA7701_14

*Elhai, J. D., Gold, S. N., Sellers, A. H., & Dorfman, W. I. (2001b). The detection of malingered posttraumatic stress disorder with MMPI-2 fake bad indices. *Assessment, 8*(2), 221-236. https://doi.org/10.1177/107319110100800210

Fariña, F., Arce, R., & Novo, M. (2002). Heurístico de anclaje en las decisiones judiciales. *Psicothema, 14*(1), 39-46. http://www.psicothema.com/pdf/684.pdf

Fariña, F., Arce, R., & Real, S. (1994). Ruedas de identificación: De la simulación y la realidad [Lineups: A comparision of high fidelity research and research in a real context]. *Psicothema, 6*(3), 395-402. https://www.psicothema.com/pdf/935.pdf

Fariña, F., Arce, R., Vilariño, M., & Novo, M. (2014). Assessment of the standard forensic procedure for the evaluation of psychological injury in intimate-partner violence. *Spanish Journal of Psychology, 17*, E32. https://doi.org/10.1017/sjp.2014.30

Fariña, F., Redondo, L., Seijo, D., Novo, M., & Arce, R. (2017). A meta-analytic review of the MMPI validity scales and indexes to detect defensiveness in custody evaluations. *International Journal of Clinical and Health Psychology, 17*(2), 128-138. http://doi.org/10.1016/j.ijchp.2017.02.002

Fox, D. D., Gerson, A., & Lees-Haley, P. R. (1995). Interrelationship of MMPI-2 validity scales in personal injury claims. *Journal of Clinical Psychology, 51*(1), 42-47. https://doi.org/10.1002/1097-4679(199501)51:1<42::AID-JCLP2270510108>3.0.CO;2-Z

Gancedo, Y., Novo, M., & Arce, R. (2020). ¿Difiere la capacidad de clasificación de respuestas simuladas de la escala de F del MMPI2 cuando se contrasta con un grupo control y con la población normativa? Una revisión meta-analítica [Does the performance to classify malingering responses of MMPI-2 F scale differ when is contrasted with a control group and with the normative sample? A meta-analytic review]. In A. M. Martín, F. Fariña, F., & R. Arce (Eds.), *Psicología jurídica y forense: Investigación para la práctica profesional*. *Colección Psicología y Ley, 16* (pp. 331-353). Walter de Gruyter. https://doi.org/10.2478/9788395609596-024

Gancedo, Y., Sanmarco, J., Marcos, V., & Seijo, M. D. (2021). Estimating the efficacy of Fptsd Scale to report malingering of PTSD: A meta-analytic review. *Revista Iberoamericana de Psicología y Salud, 12*(1), 44-57. https://doi.org/10.23923/j.rips.2021.01.044

Gancedo, Y., Sanmarco, J., Seijo, D., & Fariña, F. (2020). Malingering evaluation: A contrastive meta-analytic review of F and F-r scales. In M. Novo & D. Seijo (Eds.), *Psychology and Law: Research for practice* (pp. 99-123). Walter de Gruyter. https://doi.org/10.2478/9788395669682-009

Gancedo, Y., Selaya, A., & Arce, R. (2020). ¿Discrimina la escala de Ds del MMPI la simulación de respuestas honestas y de pacientes clínicos? Una revisión meta-analítica [Does the MMPI Ds scale discriminate between malingering responses from honest and real patient responses? A meta-analytic review]. In A. M. Martín, F. Fariña, F., & R. Arce (Eds.), *Psicología jurídica y forense: Investigación para la práctica profesional*. *Colección Psicología y Ley, 16* (pp. 11-32). Walter de Gruyter. https://doi.org/10.2478/9788395609596-001

García, C. A., & Romero, J. C. (2009). Comparación de las propiedades psicométricas del MMPI y el MMPI-2 en una muestra de universitarios bogotanos. *Revista Colombiana de Psicología, 18*(1), 19-31. https://revistas.unal.edu.co/index.php/psicologia/article/view/1607/10843

*Gassen, M. D., Pietz, C. A., Spray, B. J., & Denney, R. L. (2007). Accuracy of Megargee's Criminal Offender Infrequency (Fc) Scale in detecting malingering among forensic examinees. *Criminal Justice and Behavior, 34*(4), 493-504. https://doi.org/10.1177/0093854806295859

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5*, 351-379. https://www.jstor.org/stable/1167179?seq=1

Gough, H. G. (1947). Simulated patterns on the Minnesota Multiphasic Personality Inventory. *Journal of Abnormal and Social Psychology, 42*, 215-225. https://doi.org/10.1037/h0063295

Gough, H. G. (1950). The F minus K dissimulation index on the Minnesota Multiphasic Personality Inventory. *Journal Consulting Psychology, 14*, 408-413. https://doi.org/10.1037/h0054506

Graham, J. R. (1990). *MMPI-2: Assessing personality and psychopathology*. Oxford University.

Graham, J. R. (2011). *MMPI-2: Assessing personality and psychopathology* (5th ed.). Oxford University Press.

Greene, R. L. (2011). *The MMPI-2/MMPI-2-RF: An interpretive manual* (3th ed.). Allyn and Bacon.

*Greiffenstein, M. F., Baker, W. J., Axelrod, B., Peck, E. A., & Gervais, R. (2004). The fake bad scale and MMPI-2 F-Family in detection of implausible psychological trauma claims. *Clinical Neuropsychologist, 18*(4), 573-590. https://doi.org/10.1080/13854040490888512

*Greiffenstein, M. F., Baker, W. J., Gola, T., Donders, J., & Miller, L. (2002). The Fake Bad Scale in a typical and severe closed head injury litigants. *Journal of Clinical Psychology, 58*(12), 1591-1600. https://doi.org/10.1002/jclp.10077

*Greiffenstein, M. F., Gola, T., & Baker, W. J. (1995). MMPI-2 validity scales versus domain specific measures in detection of factitious traumatic brain injury. *Clinical Neuropsychologist, 9*(3), 230-240. https://doi.org/10.1080/13854049508400485

*Greve, K. W., Bianchini, K. J., Love, J. M., Brennan, A., & Heinly, M. T. (2006). Sensitivity and specificity of MMPI-2 validity scales and indicators to malingered neurocognitive dysfunction in traumatic brain injury. *Clinical Neuropsychologist, 20*(3), 491-512. https://doi.org/10.1080/13854040590967144

Hawkins, D. M. (1980). *Identification of outliers.* Springer.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128. https://doi.org/10.3102/10769986006002107

Henry, G. K., Heilbronner, R. L., Mittenberg, W., Enders, C., & Roberts, D. C. (2008). Empirical derivation of a new MMPI-2 scale for identifying probable malingering in personal injury litigants and disability claimants: The 15-item Malingered Mood Disorder Scale (MMDS). *Clinical Neuropsychologist, 22*, 158-168. https://doi.org/10.1080/13825580601025916

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Sage.

*Iverson, G. L., Franzen, M. D., & Hammond, J. A. (1995). Examination of inmates' ability to malinger on the MMPI-2. *Psychological Assessment, 7*(1), 118-121. https://doi.org/10.1037/1040-3590.7.1.118

*Jana, Y. A. (2001). *The effectiveness of the MMPI-2 in detecting malingered schizophrenia in adult female inmates in Puerto Rico who receive coaching on diagnostic-specific criteria* [Doctoral dissertation, California School of Professional Psychology]. Los Angeles ProQuest Dissertation & Theses. https://www.proquest.com/docview/251284639?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 617-627. https://doi.org/10.1001/archpsyc.62.6.617

Kessler, R. C., Sonnega, A., Hughes, M., & Nelson, C. B. (1995). Posttraumatic stress disorder in the National Comorbidity Survey. *Archives of General Psychiatry, 52*, 1048-1060. https://doi.org/10.1001/archpsyc.1995.03950240066012

*Kopf, T., Galić, S., & Matešić, K. (2016). The efficiency of MMPI-2 validity scales in detecting malingering of mixed anxiety-depressive disorder. *Alcoholism and Psychiatry Research, Journal on Psychiatric Research and Addictions, 52*(1), 33-50. https://hrcak.srce.hr/en/160022

*Kucharski, L. T., & Johnsen, D. (2002). A comparison of simulation and known groups in the detection of malingering on the MMPI-2. *Journal of Forensic Sciences, 47*(5), 1078-1082. https://doi.org/10.1520/JFS15512J

*Kurtz, R. A. (1992). *The vulnerability of the MMPI-2, M Test, and SIRS to two strategies of malingering psychosis in a forensic setting* [Doctoral dissertation, University of Louisville]. ProQuest Dissertations Publishing. https://www.proquest.com/docview/303998039?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses

*Lange, R. T., Sullivan, K. A., & Scott, C. (2010). Comparison of MMPI-2 and PAI validity indicators to detect feigned depression and PTSD symptom reporting. *Psychiatry Research, 176*(2-3), 229-235. https://doi.org/10.1016/j.psychres.2009.03.004

*Larrabee, G. J. (2003). Detection of symptom exaggeration with the MMPI-2 in litigants with malingered neurocognitive dysfunction. *Clinical Neuropsychologist, 17*(1), 54-68. https://doi.org/10.1076/clin.17.1.54.15627

Lees-Haley, P. R. (1989). MMPI F and F–K scales: Questionable indices of malingering. *American Journal of Forensic Psychology, 7*(1), 81-84.

Lees-Haley, P. R. (1991). MMPI-2 F and F–K scores of personal injury malingerers in vocational neuropsychological and emotional distress claims. *American Journal of Forensic Psychology, 9*(3), 5-14.

*Lees-Haley, P. R. (1992). Efficacy of MMPI-2 validity scales and MCMI-II modifier scales for detecting spurious PTSD claims: F, F–K, Fake Bad Scale, Ego Strength, Subtle-Obvious subscales, DIS, and DEB. *Journal of Clinical Psychology, 48*(5), 681-689. https://doi.org/10.1002/1097-4679(199209)48:5<681::AID-JCLP2270480516>3.0.CO;2-Q

Lees-Haley, P. R., Iverson, G. L., Lange, R. T., Fox, D. D., & Allen III, L. M. (2002). Malingering in forensic neuropsychology: Daubert and the MMPI-2. *Journal of Forensic Neuropsychology, 3*(1-2), 167-203. https://doi.org/10.1300/J151v03n01_01

*Lewis, J. L., Simcox, A. M., & Berry, D. T. R. (2002). Screening for feigned psychiatric symptoms in a forensic sample by using the MMPI-2 and the Structured Inventory of Malingered Symptomatology. *Psychological Assessment, 14*(2), 170-176. https://doi.org/10.1037/1040-3590.14.2.170

Merckelbach, H., Smeets, T., & Jelicic, M. (2009). Experimental simulation: Type of malingering scenario makes a difference. *Journal of Forensic Psychiatry & Psychology, 20*(3), 378-386. https://doi.org/10.1080/14789940802456686

*Mihura, J. L., Schlottmann, R. S., & Scott, A. B. (2000). Are the MMPI subtle subscales subtle measures of their scales? *Journal of Clinical Psychology, 56*(1), 139-148. https://doi.org/10.1002/(SICI)1097-4679(200001)56:1<139::AID-JCLP13>3.0.CO;2-1

*Moayedi, N. (2013). *The assessment of malingering using the MMPI-2* [Doctoral dissertation, Alliant International University]. Dissertation Abstracts International: Section B: The Sciences and Engineering. https://www.proquest.com/docview/1317633593?sourcetype=Dissertations%20&%20Theses

Montes, A., Sanmarco, J., Novo, M., Cea, B., & Arce, R. (2022). Estimating the psychological harm consequence of bullying victimization: A meta-analytic review for forensic evaluation. *International Journal of Environmental Research and Public Health, 19*, 13852. https://doi.org/10.3390/ijerph192113852

*Moran, P. J. (1999). *MMPI-2 profiles of students instructed to overreport depressive and manic psychopathology using either specific symptom information or symptom stereotypes* [Doctoral dissertation, Pacific Graduate School of Psychology]. https://www.proquest.com/docview/304558190?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses

Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika, 8*, 161-168. https://doi.org/10.1007/BF02288700

Mundia, L. (2011). Social desirability, non-response bias and reliability in a long self-report measure: illustrations from the MMPI-2 administered to Brunei student teachers. *Educational Psychology, 31*(2), 207–224. https://doi.org/10.1080/01443410.2010.545049

Nelson, N. W., Sweet, J. J., & Demakis, G. J. (2006). Meta-Analysis of the MMPI-2 Fake Bad Scale: Utility in Forensic Practice. *Clinical Neuropsychologist, 20*(1), 39-58. https://doi.org/10.1080/13854040500459322

Nelson, N. W., Sweet, J. J., & Demakis, G. J. (2010). Updated Meta-analysis of the MMPI-2 Fake Bad Scale: Verified Utility in Forensic Practice. *Clinical Neuropsychologist, 20*(1), 39-58. https://doi.org/10.1080/13854040500459322

Osborne, D., Colligan, R., & Offord, K. (1986). Normative tables for the F–K index of the MMPI based on a contemporary normal sample. *Journal of Clinical Psychology, 42*(4), 593-595. https://doi.org/10.1002/1097-4679(198607)42:4<593::aid-jclp2270420408>3.0.co;2-n

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

*Poggioli, R. A. (2000). *Malingered paranoid schizophrenia on the MMPI-2: Effects of coaching and prior exposure to schizophrenic inpatients* [Doctoral dissertation, Hofstra University]. ProQuest Dissertations Publishing. https://www.proquest.com/docview/304591875

Pope, K. S., Butcher, J. N., & Seelen, J. (2006). The MMPI, MMPI-2, and MMPI-A in court testimony. In K. S. Pope, J. N. Butcher, & J. Seelen, *The MMPI, MMPI-2, & MMPI-A in court: A practical guide for expert witnesses and attorneys* (3th ed., pp. 7-36). American Psychological Association. https://doi.org/10.1037/11437-002

Puente-López, E., Pina, D., López-López, R., González, H., Bošković, I., & Merten, T. (2023). Prevalence estimates of symptom feigning and malingering in Spain. *Psychological Injury and Law, 16*(1), 1-17. https://doi.org/10.1007/s12207-022-09458-w

Puente-López, E., Pina, D., Daugherty, J.C., Pérez-García, M., y Merten, T. (2024). Simulación y validez de la información de síntomas psicopatológicos en España: conceptos, métodos y desafíos [Malingering and validity of reported psychopathological symptoms in Spain: concepts, methods and challenges]. *Revista Iberoamericana de Psicología y Salud, 15*(2), 66-79. https://doi.org/10.23923/j.rips.2024.02.077

Redondo, L., Fariña, F., Seijo, D., Novo, M., & Arce, R. (2019). A meta-analytical review of the responses in the MMPI-2/MMPI-2-RF clinical and restructured scales of parents in child custody dispute. *Anales de Psicología, 35*(1), 156-165. https://doi.org/10.6018/analesps.35.1.338381

Rogers, R. (2018a). Detection strategies for malingering and defensiveness. In R. Rogers & S. D. Bender (Eds.), *Clinical assessment of malingering and deception* (4th ed., pp. 18-41). Guilford Press.

Rogers, R. (2018b). Structured interviews and dissimulation. In R. Rogers & S. D. Bender (Eds.), *Clinical assessment of malingering and deception* (4th ed., pp. 301-322). Guilford Press.

Rogers, R. (2018c). Researching response styles. In R. Rogers & S. D. Bender (Eds.), *Clinical assessment of malingering and deception* (4th ed., pp. 592-614). Guilford Press.

Rogers, R. Sewell, K. W., Martin, M. A., & Vitacco, M. J. (2003). Detection of feigned mental disorders: A meta-analysis of the MMPI-2 and malingering. *Assessment, 10*(2), 160-177. https://doi.org/10.1177/1073191103010002007

Rogers, R., Donnelly, J. W., & Correa, A. A. (2019). Translated measures in forensic evaluations with specific applications to feigned mental disorders. *Psychological Injury and Law, 12*(3-4), 191-203. https://doi.org/10.1007/s12207-019-09362-w

Rogers, R., Sewell, K. W., & Salekin, R. T. (1994). A meta-analysis of malingering on the MMPI-2. *Assessment, 1*(3), 227-237. https://doi.org/10.1177/107319119400100302

*Rogers, R., Sewell, K. W., & Ustad, K. L. (1995). Feigning among chronic outpatients on the MMPI-2: A systematic examination of fake-bad indicators. *Assessment, 2*(1), 81-89. https://doi.org/10.1177/1073191195002001008

*Ross, S. R., Millis, S. R., Krukowski, R. A., Putnam, S. H., & Adams, K. M. (2004). Detecting incomplete effort on the MMPI-2: An examination of the fake-bad scale in mild head injury. *Journal of Clinical and Experimental Neuropsychology, 26*(1), 115-124. https://doi.org/10.1076/jcen.26.1.115.23933

*Sánchez, G., Jiménez, F., & Ampudia, A. (2008). Detectando el perfil simulador en el MMPI-2: Una propuesta basada en la investigación. *Revista de Psicología, 26*(2), 277-298. https://doi.org/10.18800/psico.200802.004

*Sánchez, G., Jiménez, F., Merino, V., & Ampudia, A. (2007). La contribución de la escala de Inconsistencia de Respuesta (IR) de Sewell y Rogers (1994) a la validez del MMPI-2. *Revista de Psicología General y Aplicada, 60*(1), 103-118. http://hdl.handle.net/10366/55786

*Schaugaard, M. J. (1999). Detection of MMPI-2 faked, honest response, and archival comparison group membership. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 60*(3-B), 1314. https://www.proquest.com/docview/304538902?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3th ed.). Sage. https://doi.org/10.4135/9781483398105

Sentencia del Tribunal Supremo 1029/1997 (Sala de lo Penal, Sección 2a), de 29 de diciembre de 1997. https://vlex.es/vid/57621765

Sentencia del Tribunal Supremo 213/2002 (Sala de lo Penal, Sección 2a), de 14 de febrero de 2002. https://vlex.es/vid/agresion-sexual-absolucion-fa-u-15055496

Sharf, A. J., Rogers, R., Williams, M. M., & Henry, S. A. (2017). The effectiveness of the MMPI-2-RF in detecting feigned mental disorders and cognitive deficits: A meta-analysis. *Journal of Psychopathology and Behavioral Assessment, 39*(3), 441-455. https://doi.org/10.1007/s10862-017-9590-1

Silberman, S. A. (2002). *The validity of the MMPI-2 in a Spanish-speaking population* [Doctoral dissertation, Fairleigh Dickinson University]. https://www.proquest.com/docview/305511312?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses

*Sivec, H. I., Lynn, S. J., & Garske, J. P. (1994). The effect of somatoform disorder and paranoid psychotic role-related dissimulations as a response set on the MMPI-2. *Assessment, 1*(1), 69-82. https://doi.org/10.1177/1073191194001001010

*Steffan, J. S., Clopton, J. R., & Morgan, R. D. (2003). An MMPI-2 Scale to Detect Malingered Depression (Md Scale). *Assessment, 10*(4), 382-392. https://doi.org/10.1177/1073191103259548

*Steffan, J. S., Morgan, R. D., Lee, J., & Sellbom, M. (2010). A comparative analysis of MMPI-2 malingering detection models among inmates.

Assessment, 17(2), 185-196. https://doi.org/10.1177/1073191109359382

*Storm, J., & Graham, J. R. (2000). Detection of coached general malingering on the MMPI-2. *Psychological Assessment, 12*(2), 158-165. https://doi.org/10.1037/1040-3590.12.2.158

*Sweet, J. J., Malina, A., & Ecklund-Johnson, E. (2006). Application of the New MMPI-2 Malingered Depression Scale to individuals undergoing neuropsychological evaluation: Relative lack of relationship to secondary gain and failure on validity indices. *Clinical Neuropsychologist, 20*(3), 541-551. https://doi.org/10.1080/13854040590967135

Tarescavage, A. M., Alosco, M. L., Ben-Porath, Y. S., Wood, A., & Luna-Jones, L. (2015). Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF) scores generated from the MMPI-2 and MMPI-2-RF test booklets: Internal structure comparability in a sample of criminal defendants. *Assessment, 22*(2), 188-197. https://doi.org/10.1177/1073191114537347

United Nations [UN]. (1985). *Declaration of basic principles of justice for victims of crime and abuse of power*. United Nations. https://www.ohchr.org/en/instruments-mechanisms/onstruments/declaration-basic-principles-justice-victims-crime-and-abuse

*Vaughan, A. E. (1995). *Detecting malingering among federal inmates using the MMPI-2* [Doctoral dissertation, Hofstra University]. ProQuest Information & Learning. https://www.proquest.com/docview/304211252/fulltextPDF/3FA6CEA1D9134077PQ/1?accountid=17253&sourcetype=Dissertations%20&%20Theses

*Viglione, D. J., Wright, D. M., Dizon, N. T., Moynihan, J. E., DuPuis, S., & Pizitz, T. D. (2001). Evading detection on the MMPI-2: Does caution produce more realistic patterns of responding? *Assessment, 8*(3), 237-250. https://doi.org/10.1177/107319110100800301

Vilariño, M., Amado, B. G., & Martín-Peña, J. (2020). Simulación de psicopatología en la incapacidad temporal: utilidad forense del SCL-90-R. In A. M. Martín, F. Fariña, & R. Arce (Eds.), *Psicología jurídica y forense: Investigación para la práctica profesional* (pp. 25-40). Sociedad Española de Psicología Jurídica y Forense.

Vilariño, M., Arce, R., & Fariña, F. (2013). Forensic-clinical interview: Reliability and validity for the evaluation of psychological injury. *European Journal of Psychology Applied to Legal Context, 5*(1), 1-21. https://journals.copmadrid.org/ejpalc/art/8c1b6fa97c4288a4514365198566c6fa

*Walters, G. L. (1998). *The effect of symptom information and validity scale information on the malingering of depression on the MMPI-2* [Doctoral dissertation, Texas Tech University]. Electronic Theses and Dissertations. http://hdl.handle.net/2346/9287

*Walters, G. L., & Clopton, J. R. (2000). Effect of symptom information and validity scale information on the malingering of depression on the MMPI-2. *Journal of Personality Assessment, 75*(2), 183-199. https://doi.org/10.1207/S15327752JPA7502_1

*Wetter, M. W., & Deitsch, S. E. (1996). Faking specific disorders and temporal response consistency on the MMPI-2. *Psychological Assessment, 8*(1), 39-47. https://doi.org/10.1037/1040-3590.8.1.39

*Wetter, M. W., Baer, R. A., Berry, D. T. R., & Reynolds, S. K. (1994). The effect of symptom information on faking on the MMPI-2. *Assessment, 1*(2), 199-207. https://doi.org/10.1177/1073191194001002010

*Wetter, M. W., Baer, R. A., Berry, D. T. R., Robison, L. H., & Sumpter, J. (1993). MMPI-2 profiles of motivated fakers given specific symptom information: A comparison to matched patients. *Psychological Assessment, 5*(3), 317–323. https://doi.org/10.1037/1040-3590.5.3.317

*Wetter, M. W., Baer, R. A., Berry, D. T. R., Smith, G. T., & Larsen, L. H. (1992). Sensitivity of MMPI-2 validity scales to random responding and malingering. *Psychological Assessment, 4*(3), 369-374. https://doi.org/10.1037/1040-3590.4.3.369

Funding: This research has been partially sponsored by a grant of the Ministry of Science and Innovation of Spain (Code: PID2020-115881RB-I00).

Conflict of interests: The authors declare no competing interests.

Data Availability: The raw data supporting the conclusions of this article are available on Annex.

Informed Consent Statement: Not applicable.

Institutional Review Board Statement: Not applicable.

Annex. Primary studies data

Reference	Source	Design	Simulation group		Control group		d^1	d^2
			N	Participants	*N*	Participants		
Austin (1992)	Paper	SR	37	Students	33	Students	4.43	4.62
Bagby et al. (1994)	Paper	SR	58	Students	90	Students	2.87	3.16
Bagby et al. (2000)	Paper	SR	23	Health professionals	---	---	---	1.75
Bagby et al. (1997a)	Paper	SR	40	Students	40	Students	3.17	3.13
Bagby et al. (1997b)	Paper	SR	78	Students	---	---	---	2.53
				Psychiatric patients	---	---	---	
				Psychologists	---	---	---	
Berry et al. (1995)	Paper	SR	18	Non-inpatients	20	General population	1.49	1.86
Berry et al. (1996)	Paper	SR	30	Inpatients	---	---	---	3.80
Berry et al. (2001)	Paper	SR	59	Non-inpatients	---	---	---	0.15
Bianchini et al. (2008)	Paper	KGC	58	Students	---	---	---	0.63
				Non-inpatients	---	---	---	
Blanchard et al. (2003)	Paper	SR	52	Students	---	---	---	4.00
Bury & Bagby (2002)	Paper	SR	131	Students	---	---	---	1.58
Cavenagh (2008)	Doctoral dissertation	SR	30	Litigants	---	---	---	0.97
Chang et al. (2017)	Paper	SR	40	Students	30	Students	4.63	4.52
Charles (1999)	Doctoral dissertation	KGC	42	Litigants	53	Honest litigants	1.27	0.75
Cramer (1995)	Paper	SR	124	Students	31	Students	1.89	---
Crawford et al. (2006)	Paper	SR	27	Students	---	---	---	0.88
Elhai et al. (2001a)	Paper	SR	79	Students	---	---	---	3.07
Elhai et al. (2001b)	Paper	SR	80	Students	---	---	---	3.07
Gassen et al. (2007)	Paper	KGC	29	Prison inmates	37	Prison inmates	2.86	2.89
Greiffenstein et al. (1995)	Paper	KGC	121	Non-inpatients	---	---	---	0.17
Greiffenstein et al. (2002)	Paper	KGC	159	Litigants	---	---	---	0.13
Greiffenstein et al. (2004)	Paper	DPD	89	Non-inpatients	---	---	---	0.33
Greve et al. (2006)	Paper	KGC	162	Non-inpatients	---	---	---	0.37
Iverson et al. (1995)	Paper	SR	28	Prison inmates	---	---	---	2.95
Jana (2001)	Doctoral dissertation	SR	33	Prison inmates	15	Prison inmates	0.62	2.21
Kopf et al. (2016)	Paper	SR	47	Students	---	---	---	3.49
Kucharski & Johnsen (2002)	Paper	SR	90	Prison inmates	30	Prison inmates	2.09	2.74
Kurtz (1992)	Doctoral dissertation	SR	120	Prison inmates	20	Prison inmates	4.37	3.09
Lange et al. (2010)	Paper	SR	29	Students	20	Students	2.29	2.24
Larrabee (2003)	Paper	KGC	26	Litigants	---	---	---	-0.09
Lees-Haley (1992)	Paper	KGC	55	Litigants	64	Honest litigants	2.445	0.92
Lewis et al. (2002)	Paper	KGC	24	Litigants	---	---	---	2.67
Mihura et al. (2000)	Paper	SR	20	Students	20	Students	-4.32	0.38
Moayedi (2013)	Doctoral dissertation	SR	30	General population	---	---	---	0.89
Moran (1999)	Doctoral dissertation	SR	102	Students	---	---	---	1.27
Poggioli (2000)	Doctoral dissertation	SR	120	General population	---	---	---	3.29
				Health professionals	---	---	---	
Rogers et al. (1995)	Paper	SR	42	Non-inpatients	---	---	---	3.82
Ross et al. (2004)	Paper	DPD	59	Litigants	---	---	---	0.48
Sánchez et al. (2007)	Paper	SR	267	General population	541	General population	4.47	4.32
Sánchez et al. (2008)	Paper	SR	272	General population	1723	General population	4.79	4.33
Schaugaard (1999)	Doctoral dissertation	SR	15	Students	15	Students	2.11	2.35
Sivec et al. (1994)	Paper	SR	179	Students	58	Students	1.96	2.92
Steffan et al. (2003)	Paper	SR	92	Students	---	---	---	2.21
Steffan et al. (2010)	Paper	SR	45	Prison inmates	---	---	---	2.46
Storm & Graham (2000)	Paper	SR	440	Students	---	---	---	2.67
Sweet et al. (2006)	Paper	DPD	89	Litigants	---	---	---	-0.06
Vaughan (1995)	Paper	SR	60	Prison inmates	30	Prison inmates	2.57	3.13
Viglione et al. (2001)	Paper	SR	88	Students	---	---	---	1.25
Walters (1998)	Doctoral dissertation	SR	342	Students	45	Students	1.69	1.98

Reference	Source	Design	Simulation group		Control group		d^1	d^2
			N	Participants	*N*	Participants		
Walters & Clopton (2000)	Paper	SR	370	Students	95	Students	1.78	1.98
Wetter et al. (1992)	Paper	SR	70	Students	68	Students	2.07	3.60
Wetter et al. (1993)	Paper	SR	42	General population	---	---	---	3.80
Wetter et al. (1994)	Paper	SR	46	General population	36	General population	2.28	2.41
Wetter & Deitsch (1996)	Paper	SR	80	Students	44	Students	0.32	1.84

Note. SR: simulation research; KGC: known-groups comparison; DPD: differential prevalence design; d^1: Cohen's d for the comparison of the simulation group with in-study control group; d^2: Cohen's d for the comparison of the simulation group with the combined control group.