



Psychosocial Intervention

<https://journals.copmadrid.org/pi>



Testing the Evaluation of the Testimony of the Global Evaluation System in Bullying Cases

Álvaro Montes, Blanca Cea, and Ramón Arce

Universidade de Santiago de Compostela, Spain

ARTICLE INFO

Received 5 May 2026
Accepted 14 September 2026

Keywords:

School bullying
Witness credibility
Witness memory
Criterial validity
Undeutsch hypothesis
Presumption of innocence

ABSTRACT

Background/aim: The prevalence of school bullying victimization has been estimated between 5% (strict criteria for prevalence classification) and 33% (lenient criteria for prevalence assessment) of children and adolescents. Bullying victimization entails decision-making at the administrative (school setting), civil (victim compensation), and criminal (criminal responsibility) levels, which must be based on certainty about the truthfulness of the allegation. A high-fidelity simulation study was designed with the aim of testing the evaluation of testimony of the Global Evaluation System (GES) in cases of bullying allegations. **Method:** A total of 154 adolescents aged from 13 to 18 years, 86 boys, were randomly assigned to the condition of testimony of a memory of real-life negative experiences or a fabricated memory of a bullying event. The protocols (statements) were subjected to a content analysis using the categories of validity of the evidence and the content categories of memories of real-life experiences of the GES. **Results:** The results showed excellent reliability, $\alpha = .874$, 95% CI [.843, .901], for the GES content category system. The findings also supported the association between the GES content categories and memories of real-life experiences. Conversely, no significant association was found between fabricated memories of a bullying event and the GES content categories, while a lack of reliability was observed, $\alpha = .476$, 95% CI [.289, .632], in this condition. Finally, strict decision criteria were computed (evidence-based criterial validity) that guarantees compliance with the principle of presumption of innocence (no innocent person should be convicted; type I error/false positives = 0): coding of ≥ 10 content categories in the complete model, ≥ 8 in the higher-prevalence model (content categories associated significantly with memories of real-life experiences), and ≥ 5 in the robust model (content categories associated significantly with memories of real-life experiences with a moderate or higher than moderate effect size). **Conclusions:** These findings support the validity and reliability of the GES content analysis categorical system for the validation of truthful testimonies of school bullying victimization (forensic task), providing scientific evidence (criterial validity) to ensure the strict compliance of the principle of presumption of innocence, but not for the classification of fabricated memories (investigative task).

The systematic study of school bullying (bullying here in after) began in the 1970s (Olweus, 1973) and has since produced substantial scientific evidence on its prevalence, effects, and prevention (Smith, 2016). However, discrepancies in the conceptualization of the phenomenon have led to inconsistencies in estimates of its prevalence and effects (Vivolo-Kantor et al., 2014). Based on Olweus's (1993) seminal definition, "a student is being bullied or victimized when he or she is exposed, repeatedly and over time, to negative actions on the part of one or more students" (p. 9), three diagnostic criteria can be identified. First, the intention to cause harm criterion implies the existence of violent behaviors that produce physical or psychological harm (harm criterion) and that are enacted with the purpose of harming (intentionality criterion), thereby distinguishing

bullying from other harmful behaviors among peers, such as rough play or jokes (Arce et al., 2014). Second, the repetition criterion implies that the behaviors are frequent and sustained over time, using a frequency of "2 or 3 times a month" over at least two months as the gold standard (Solberg & Olweus, 2003). Third, the power imbalance criterion refers to an asymmetrical relationship in which the victim is unable to defend themselves against the perpetrator due to physical, psychological, or social differences (Smith et al., 2012). These criteria allow for the effective identification of individuals involved in bullying and for the differential diagnosis of bullying from other forms of victimization during childhood and adolescence, such as dating violence, sexual violence, child maltreatment, physical fights, rough play, or hurtful teasing (Centers for Disease Control

Cite this article as: Montes, A., Cea, B., & Arce, R. (2026). Testing the evaluation of the testimony of the global evaluation system in bullying cases. *Psychosocial Intervention*, 35, Article e260820. <https://doi.org/10.5093/pi2026a12>

Funding: This research has been partially sponsored by + the Ministry of Science and Innovation of Spain (PID2020-115881RB-I00) and by a grant to Álvaro Montes from the Spanish Ministry of Universities under the program "Ayudas para la Formación de Profesorado Universitario" (Code: FPU20/02169).
Correspondence: ramon.arce@usc.es (R. Arce).

ISSN:1132-0559/© 2026 Colegio Oficial de la Psicología de Madrid. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

and Prevention [CDC, 2022]; United Nations Educational, Scientific and Cultural Organization [UNESCO, 2019]). Additionally, Arce et al. (2014) incorporated the criterion of victimization, defined as the verification of harm, such that, in the absence of harm, there is no victim of harassment. Harm includes physical or mental injury, emotional suffering, economic loss or substantial impairment of their fundamental rights (United Nations [UN, 1985]). Since the 2000s, initially linked to the emergence of the internet and later to the rise of social media and smartphones, literature has identified digital environments as contexts conducive to harassment, commonly referred to as cyberbullying, but also as cyber-aggression, electronic bullying, cyber harassment, electronic victimization, cyber-control or online bullying (Chun et al., 2020; Englander et al., 2017; Kwan et al., 2020; Niehaus et al., 2025). However, there is no consensus on its conceptualization. Thus, some authors include cyberbullying within the seminal concept of bullying (traditional bullying), arguing that the diagnostic criterion of power imbalance may not apply in digital environments because there is no direct contact with the victim and interactions may occur anonymously, while redefining the repetition criterion as the number of people who have accessed the message or image or the length of time it remains available (Olweus & Limber, 2018; Smith et al., 2012). Therefore, the most appropriate term would be bullying among school-aged children and adolescents, regardless of whether it occurs face to face or through technologies such as mobile phones, the internet, or social media (Thomsen et al., 2024), facilitating the differential diagnosis of bullying perpetrated by adults (Arce et al., 2014). Furthermore, evidence of overlap between traditional bullying and cyberbullying supports the view that they represent the same construct measured in different contexts, as harassment by electronic means rarely occurs in isolation (Chun et al., 2020; Kim et al., 2017; Olweus, 2012; Pichel et al., 2021). Conversely, other researchers consider cyberbullying to be an independent phenomenon with its own characteristics (Cosma et al., 2020; Kwan et al., 2020), making it difficult to distinguish from other forms of online aggression (which share characteristics with harassment) and from violent behaviors such as sexual abuse or sexual assault (Marcos, Seijo, et al., 2024). Moreover, in the legal context, harassment and other forms of violence constitute distinct criminal offences, requiring psychological assessment to establish a differential diagnosis between them (Marcos, Seijo, et al., 2024), consistent with psychology's role in assisting the justice system.

Conceptual heterogeneity and, above all, differences in assessment approaches have contributed to the wide variability in prevalence estimates. Specifically, prevalence estimates based on psychometric assessments or surveys that measure exposure to aggressive behaviors (lenient criteria) differ substantially from those based on diagnostic approaches that identify cases of bullying victimization only when the established diagnostic criteria for bullying are met (strict criteria). Accordingly, prevalence estimates range from approximately one-third of children and adolescents when lenient assessment criteria are used (Modecki et al., 2014; UNESCO, 2019) to above 5% when strict diagnostic criteria are applied (Arce et al., 2014; Díaz-Aguado et al., 2024).

The mental health consequences of bullying victimization have been well documented, including both internalizing mental health problems, such as depression, anxiety, psychosomatic symptoms and suicidal ideation (Katsaras et al., 2018; Li et al., 2024; Marcos, Montes, et al., 2024; Montes et al., 2022; Moore et al., 2017; Ye et al., 2023), and externalizing mental health problems, such as aggressive behaviors, hyperactivity, and addictive behaviors (Ak et al., 2015; Bradshaw et al., 2015; Kowalski et al., 2014; Moore et al., 2017; Pichel et al., 2022; Rial-Boubeta et al., 2024; Woo et al., 2019). In the academic setting, adverse effects have been observed in academic failure, higher absenteeism and lower sense of community (Davis et al., 2019; Garaigordobil et al., 2019; Hong et al., 2018; Laith & Vaillacourt, 2022; Lucas-Molina et al., 2022; Marcos, Montes, et al., 2024; Novo

et al., 2016). Bullying victimization has also been associated with psychological harm of legal relevance, particularly Posttraumatic Stress Disorder (PTSD; American Psychiatric Association [APA, 2022]), enabling the establishment of the causal nexus between bullying victimization and psychological harm. Nonetheless, some authors (Idsoe et al., 2021; Nielsen et al., 2015) have questioned whether bullying should be considered a traumatic event. In either case, this concern does not alter the nature of the harm, nor its assessment, which would instead be classified as Adjustment Disorder when bullying does not imply a credible threat of serious harm or sexual violence (differential diagnosis of PTSD; APA, 2022). Similarly, the development of internalizing mental health symptoms in children younger than 13 years has also been questioned, with some authors suggesting that these effects may not be reliably assessed at these ages (Cea et al., 2025; Danzi & La Greca, 2016; van der Kolk et al., 2019). A meta-analytical review (Montes et al., 2022) found a significant and moderate association between bullying victimization and psychological harm ($r = .297$), while also showing that 26.7% of bullied children and adolescents did not develop psychological harm.

In this regard, the United Nations (UN, 2018), in Resolution A/73/265, urged Member States to address bullying through administrative, civil, and criminal proceedings, including victim compensation and the determination of criminal responsibility. The burden of proof (i.e., the need to prove the case) lies with the prosecution in criminal proceedings and with the claimant in civil proceedings. The central evidence in such cases consists of the evaluation of the victim's testimony and psychological harm. The victim's testimony may be sufficient to convict the accused when it is corroborated by a forensic evaluation. For this purpose, verbal content analysis of the testimony (i.e., statement) based on the Undeutsch's (1967) hypothesis, has received substantial empirical support and has been admitted in courts (Arce, 2017; Oberlader et al., 2016) across different types of witnesses (Vrij et al., 2025). Among the available procedures, the Statement Validity Assessment (SVA; Steller, 1989), which includes the content analysis technique Criteria-Based Content Analysis (CBCA; Steller & Köhnken, 1989), is particularly noteworthy. The CBCA reality criteria have been shown to validly discriminate between statements based on memories of real experiences and fabricated memories of an event (Amado et al., 2015; Oberlader et al., 2016; Wachendörfer & Oeberst, 2024). However, the CBCA categorical system does not fully meet the criteria required of forensic evidence because it lacks reliability and validity, so it is not a scientific instrument (on this regard, it has been described by Steller and Köhnken (1989) as a semi-standardized procedure); and it lacks a standardized application manual and objective decision criteria (criterial validity), thereby compromising its judicial validity (Arce, 2017). Likewise, it relies on a single interview during the assessment process (Steller, 1989) and, therefore, does not evaluate the judicial criterion of persistence in the incrimination (Arce, 2017). To overcome these limitations, Arce and Fariña (2005, 2006, 2014, 2015) developed the Global Evaluation System (GES), which enables a systematic evaluation of both testimony and psychological harm. The GES categorical system for the content analysis of memories of real-life experiences has been tested and validated in memories of victims of gender-based violence (Pérez, 2024; Vilarriño et al., 2011), sexual assaults against adults (Arce, 2024), child sexual abuse (Abeledo, 2024) and mobbing (workplace bullying; Arce & Fariña, 2011). It has also been shown to validly discriminate memories of lived experiences from memories of events that were observed but not personally experienced (Sanmarco et al., 2023). This forensic technique has solid scientific support (Castro-Gallego et al., 2024; N. Martínez & Orihuela, 2020; Mena, 2021), meets the Daubert standard (Daubert vs Merrell Dow Pharmaceuticals, 1993) and satisfies the legal and judicial requirements for forensic evidence (Arce, 2017). Furthermore, it has been validated in numerous judicial decisions involving a wide range of offences and populations (e.g., Sentencia

10554/2014 del Tribunal Supremo [sentence of the Spanish Supreme Court]). However, the GES content category system has not yet been tested in cases of bullying victimization.

In this context, the aim of the present study was to test the GES content category system for memories of real-life experiences in bullying cases. Specifically, the study examined the capacity of the GES content categories to discriminate between fabricated memories of bullying and memories of lived experiences, estimated their reliability and validity, and identified strict decision criteria (criterial validity) to minimize type I error (i.e., false positives: classification of a fabricated memory as a memory of a lived event).

Method

Participants

A total of 154 adolescents aged 13 to 18 years ($M = 15.12$, $SD = 1.01$), with a balanced gender distribution (55.8% boys and 44.2% girls), $\chi^2(1) = 2.10$, $p = .147$, participated in the study. In the sample, 89.0% ($n = 137$) were enrolled in Secondary Education (equivalent to Grades 7–10), 4.5% ($n = 7$) in Upper Secondary Education (equivalent to Grades 11–12), and 6.5% ($n = 10$) in vocational education and training.

Design and Procedure

A simulation study (Rogers, 2018) was conducted with two conditions: memories of real-life experiences vs. fabricated memories of bullying victimization. Because the main forensic challenge is determining whether a fabricated allegation can be misclassified as a statement based on a lived experience, the experimental condition was the fabricated bullying account. Participants were recruited through a non-probability convenience sampling method. No prior exclusion criteria were applied, as participant availability was the primary consideration. Participants were randomly assigned to one of the two study conditions (77 per condition). All participants received an information sheet detailing the objectives, tasks and conditions of the study, and their participation was encouraged by highlighting the relevance of the results for the assessment of bullying cases.

Participants in the autobiographical (memory of real-life experiences) condition were asked to narrate a relevant lived event, defined as a directly experienced negative situation involving a loss of control over its course and outcomes, such as adverse experiences, traffic crashes, accidents with significant consequences, animal attacks or the illness/death of a close relative (Steller, 1989). These were designed to ensure alignment with the task and to achieve a high-fidelity research design, thereby enhancing external and ecological validity and minimising the limitations of face validity associated with non-real or low-fidelity research (Konecni & Ebbesen, 1992), whose results differ significantly from those obtained in field experiments (Fariña et al., 1994). Compliance with the instructions was assessed by checking the characteristics of the research design (i.e., negative accounts, directly experienced and involving loss of control over the situation and outcomes) and the plausibility and realism of the account.

Participants assigned to the fabricated bullying victimization condition were instructed to provide a fictitious memory account (false allegation) of a bullying victimization (traditional and/or cyberbullying). To this end, they were asked to adopt the perspective of someone who simulates being a victim of bullying for a self-justified reason (e.g., change of school, revenge, concealment of academic failure) and who files a report based on fabricated events. To ensure that the account was realistic and plausible, participants were provided with guidance on bullying dynamics in the school context and were given one to two weeks to prepare their statement.

The instructions were written clearly and comprehensibly and were specific and contextualized to the study conditions. During the assessment, participants' understanding and compliance were verified. All autobiographical memory accounts were plausible, realistic, and in accordance with the instructions. In the fabricated memory condition, recall and understanding of the instructions, motivation, and preparation of the account were verified through a debriefing, with the aim of excluding those who did not remember or did not correctly follow them (Arce et al., 2015; Fariña et al., 2014). Additionally, the bullying behaviours included in the accounts were analyzed using the checklist of the UPF-4 bullying scale, to exclude those that did not report having been subjected to bullying strategies (at least two) and repeated over time (Arce et al., 2014). All fabricated accounts included more than two bullying strategies repeatedly directed at the target.

Ethical Issues

Adolescents participated voluntarily in the study and provided written informed consent; parents/legal guardians and school authorities also approved the evaluations. Data were stored and processed in accordance with Spanish Data Protection Law (Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales). The study was approved by the Research Ethics Committee of the Universidad de Santiago de Compostela (Code: USC-54/2022).

Measures

To avoid the inclusion of students who had experienced bullying during their school years and to prevent secondary victimization (Ceci et al., 2007), the UPF-4 bullying scale was applied as a screening measure, which assesses strategies of traditional bullying and cyberbullying (Arce et al., 2014). For the diagnostic impression of bullying, exposure to ≥ 3 bullying strategies repeatedly (≥ 2 times per month during at least the last two months in traditional bullying; and time and/or population audience size [public exposure to a message or image] in cyberbullying) was required.

The protocols (i.e., statements) were obtained following the standards for forensic evidence, that is, without contaminating the witness's memory with externally sourced (i.e., interviewer) or implanted post-event information (Davis & Loftus, 2007; Giorgianni et al., 2025), while ensuring the elicitation of rich and detailed memories of the event (Arce et al., 2023). Narrative interviews promote longer and less contaminated accounts than other interview techniques (Dodier & Otgaar, 2019; Fisher & Geiselman, 2010; Palena et al., 2024), thereby increasing the likelihood of obtaining sufficient evidence. Accordingly, the abbreviated version of the Cognitive Interview (CI; Fisher et al., 1989; Fisher & Geiselman, 1992), comprising mental reinstatement of context and free recall, was employed, yielding levels of productivity comparable to those of the full CI (Davis et al., 2005), while reducing external contamination (Arce, 2017; Memon et al., 2010). In addition, the abbreviated CI has been associated with greater elicitation of content criteria (Colwell et al., 2002). The CI has also demonstrated the capacity to reduce the negative consequences of recalling traumatic events, potentially increasing well-being during the process (Dodier & Otgaar, 2019; Fisher & Geiselman, 2010), particularly when rapport with the interviewee is well established. Moreover, it can be applied in repeated measures (assessment of the validity of the evidence) without affecting the content of the accounts (Deeb et al., 2024).

The interviews were conducted by two interviewers trained and experienced in using the Cognitive Interview (CI) in forensic settings. All sessions were video recorded for subsequent transcription and analysis. No significant differences were found between interviewer

1 and interviewer 2 in the number of GES content categories elicited from either autobiographical memories, $t(75) = 0.12, p = .904$, and fabricated memories of bullying victimization $t(75) = 0.96, p = .574$. In sum, the interviewers did not exert a differential effect on the elicitation of GES content categories.

Table 1. Global Evaluation System Content Categories

General Characteristics (unit of analysis: the statement as a whole)
1. Logical structure and realism in the reconstruction of the event.
2. Ellipsis (temporal and spatial gaps).
3. Amount of detail.
4. Concreteness and clarity of the account.
5. Singularity and originality of the account (non-stereotyped)
Memory Contents (unit of analysis: specific parts of the statement)
6. Contextual, spatial, and temporal embedding of the event.
7. Description of interactions (action – reaction – counterreaction)
8. Reproduction of conversations (direct or indirect speech; conversation or comment)
9. Narration of unbelievable facts ¹
10. Sensory information.
Witness Cognitions (unit of analysis: witness cognitions)
11. Cognitive operations.
12. Cognitions inconsistent with the lived experience.
13. Cognitions about affect and mental state.
14. Cognitions about the cause of victimization ¹
15. Cognitions of improper assumption of responsibility.
Content Characteristics (unit of analysis: specific parts of the statement)
16. Infrequent information regarding typical events and contexts.
17. Irrelevant information.
18. Elicitation of memory nodes or related events.
Content Regarding Memory Quality (unit of analysis: specific parts of the statement)
19. Spontaneous corrections, specifications, and elaborations of the account.
20. Admission of memory gaps.
21. Expression of doubts about one's own testimony.
22. Improper assumption of responsibility.
23. Forgiveness towards the aggressor ¹
Psychological Harm Derived from the Action under Judgment
24. Subtle clinical symptoms (not accessible to simulation) that are part of the psychological harm.

Note. ¹These content categories are not expected to be productive or mitigated the productivity in the condition of memories of real-life experiences, as they do not refer to victimization through a crime.

Analysis of Protocols

Testimony assessment using the GES comprises two measures: testimony validity (i.e., evidence validity in judicial terms) and content analysis of the statement, based on characteristic criteria of memories of real-life experiences. Testimony validity was examined through the criteria of sufficiency and validity. Sufficiency refers to whether the testimony contains enough information to be subjected to content analysis. Because fabricated memories tend to be simpler than genuine memories (Vrij et al., 2025), a sufficient statement must exceed the witness's expected memory capacity and include all necessary information about the events. Testimony validity analysis is composed of five criteria derived from a content analysis of judicial rulings (construct validity; Arce et al., 2013): internal consistency; external consistency, which was not assessed in this study because it requires comparison with other evidence in the judicial file; consistency of statements over time, referring to the stability of the core information across interviews (Montes et al., 2024); statement persistence, referring to the consistency of temporal and spatial references; and consistency with scientific and natural laws. The categorical system for assessing testimony validity showed good

internal consistency, $\alpha = .874$ (Abeledo, 2024; Arce et al., 2013; M. S. Martínez, 2024; Pérez, 2024). In the judicial context, testimony is considered valid only if all five criteria are met. To test temporal consistency between statements, two testimonies were obtained at least one week apart.

Furthermore, content analysis was conducted using the 24 GES content categories, which are applicable across all types of cases and populations. The GES content categories have shown very good internal consistency, $\alpha = .899$, and validity (construct, discriminant, predictive and criterial validity) in the classification of memories of lived experiences and in discriminating between memories of real-life experiences and fabricated memories (Abeledo, 2024; Arce et al., 2013; M. S. Martínez, 2024; Pérez, 2024; Vilariño, 2010; Vilariño et al., 2011). Although the categorical system includes two additional categories for bullying cases (minimisation of bullying and attribution of blame to the victim), these were not included in this study, as they were not productive in the condition of memory of real-life experiences. The 24 GES content categories and the unit of measurement (Arce, 2024) are presented in Table 1.

Two forensic experts with experience in content analysis using the GES coded the protocols. Both were trained using statements not included in the study. They received a manual containing the category definitions, examples, units of analysis, and decision criteria, in accordance with the Daubert standard to ensure coding fidelity. They were also instructed to code a category as present only when they were absolutely certain, as false-positive coding may lead to violations of the presumption of innocence. The protocols were assigned to the raters randomly and proportionally across both conditions. Additionally, each rater assessed approximately two thirds of the interviews analyzed by the other ($n = 54$), to estimate the between-rater concordance. Likewise, each evaluator recoded, after at least 10 days from the original coding, 20 of their protocols to estimate the within-rater concordance.

Coding Fidelity

Coding fidelity was computed using true kappa ($\bar{\kappa}$; Fariña et al., 2002), which verifies exact correspondence in the codings, while correcting for agreement by chance. The exact correspondence in the coding means that the same content must be coded into the same category; otherwise, two errors would be counted as one agreement, resulting in a false agreement. True kappa values were interpreted using the criteria of Landis and Koch (1977) were applied: slight agreement ($\bar{\kappa} \leq .20$); fair ($.21 \leq \bar{\kappa} \leq .40$); moderate ($.41 \leq \bar{\kappa} \leq .60$); substantial ($.61 \leq \bar{\kappa} \leq .80$); and almost perfect ($\bar{\kappa} \geq .81$). However, the literature has identified the kappa paradox (Zec et al., 2017), whereby the coefficient tends to underestimate agreement when coding frequencies are highly unequal (i.e., the category has a very high or very low probability of coding). Consequently, $\bar{\kappa}$ values for some categories indicate low or moderate fidelity despite high levels of agreement (see Table 2). To address this paradox, the True Agreement Index ($\overline{TAI} = \text{Agreements}/(\text{Agreements} + \text{Errors})$) was also computed, verifying exact correspondence (Arce et al., 2000). Thus, categories with $\overline{TAI} > .80$ were considered concordant (Tversky, 1977). Between-rater concordance analyses (see Table 2 for each category) showed very good overall agreement ($\bar{\kappa} = .856$, 95% CI [.821, .891]; $\overline{TAI} = .958$, 95% CI [.951, .965]); as well as for within-rater concordance ($\bar{\kappa} = .981$, 95% CI [.897, 1.00]; $\overline{TAI} = .991$, 95% CI [.983, .999]). For categories affected by the Kappa paradox, concordance was validated by the True Agreement Index, $\overline{TAI} > .80$.

Data Analysis

The productivity of the GES content categories was computed by means of the Zeta contrast between the observed probability (p_o)

Table 2. Coding Fidelity in the GES Content Categories

Content category	$\bar{\kappa}$ [95% CI]	$\bar{\text{TAI}}$ [95% CI]
Logical structure and realism	.225 [.062, .388]	.899 [.842, .956]
Ellipsis	1.00	1.00
Amount of detail	.721 [.541, .902]	.945 [.902, .988]
Concreteness and clarity of the account	.947 [.760, 1.00]	.991 [.973, 1.00]
Singularity and originality of the account	.620 [.432, .807]	.971 [.939, 1.00]
Contextual, spatial, and temporal embedding	.861 [.673, 1.00]	.963 [.927, .998]
Description of interactions	.781 [.598, .965]	.927 [.878, .976]
Reproduction of conversations	.887 [.700, 1.00]	.945 [.902, .988]
Narration of unbelievable facts	.591 [.404, .778]	.954 [.914, .994]
Sensory information	.548 [.380, .715]	.872 [.809, .935]
Cognitive operations	.790 [.600, .980]	.981 [.955, 1.00]
Cognitions inconsistent with the lived experience	—	---
Cognitions about affect and mental state	.727 [.543, .911]	.862 [.797, .927]
Cognitions about the cause of victimization	—	—
Cognitions of improper assumption of responsibility	.356 [.185, .528]	.853 [.786, .920]
Infrequent information	.732 [.551, .912]	.963 [.927, .999]
Irrelevant information	.899 [.711, 1.00]	.982 [.957, 1.00]
Elicitation of memory nodes	1.00	1.00
Spontaneous corrections	.795 [.612, .979]	.991 [.973, 1.00]
Admission of memory gaps	.486 [.318, .653]	.884 [.824, .944]
Expression of doubts about one's own testimony	.681 [.494, .868]	.954 [.914, .994]
Improper assumption of responsibility	—	—
Forgiveness towards the aggressor	—	—
Subtle clinical symptoms	1.00	1.00

Note. For some categories, consistency was not computed as they were not coded in the protocols.

and a constant, adopting the following criterion values (Fandiño et al., 2021): $p_o \leq .05$, trivial prevalence; $.05 < p_o < .50$, significant prevalence; $.50 \leq p_o < .95$, common prevalence; $p_o \geq .95$, normal prevalence. Effect size was estimated using Cohen's h , with the following benchmarks (Arce et al., 2015; Cohen, 1988): $h = 0.20$, small; $h = 0.50$, moderate; $h = 0.80$, large; and $h = 1.20$, more than large. Between-group mean differences were examined using two-sample t -test, estimating the magnitude of the effect with Cohen's d and the model error with the Probability of an Inferiority Score (PIS; Vilariño et al., 2022). The associations between each content and validity category with memory type (memories of real-life experiences vs. fabricated memories of bullying) were estimated with Fisher's exact test. Effect size was estimated using the Prevalence Ratio (PR), interpreted as small magnitude if $PR \geq 1.44$; moderate if $PR \geq 2.47$; large if $PR \geq 4.25$; and more than large if $PR \geq 8.82$, corresponding to Cohen's d values of 0.50, 0.80 and 1.20, respectively (Arce et al., 2015). Incremental validity (advantages) was assessed using the difference of proportions (p_{diff}), being interpreted as significant if the lower limit of the 95% CI is $\geq .05$ (non-significant if the lower limit of 95% CI is $< .05$); significant but non-common if the lower limit of the 95% CI is higher than .05 and the upper limit lower than .50; common if the 95% CI included .50; more than common when the lower limit is higher than .50 and the upper limit lower than .95; and general/normal if the 95% CI included .95 or the lower limit is higher than .95. When a content category was absent in the fabricated memory condition, the prevalence ratio (PR) could not be mathematically estimated. Therefore, the observed prevalence in the autobiographical memory condition is reported descriptively, whereas the corresponding PR and its 95% confidence interval are not estimable.

Likewise, as type I forensic error (i.e., false positives) has to be zero (i.e., classification of a false testimony as a true testimony), Arce (Fandiño et al., 2021; Redondo et al., 2019) developed an effect size to estimate this type of error and to create decision models: the Forensic Discrimination Index, $DI_f = (\text{true positives} +$

true negatives)/ N . This index assesses the discrimination capacity between memories of the content categories. It is interpreted as significant when the lower limit of the 95% CI of the DI_f is $> .50$ (discriminates better than chance), and non-significant when the 95% CI of DI_f includes .50 (discriminates at chance level); whereas if the upper limit is $< .50$, it discriminates worse than chance. Its magnitude is considered small if $DI_f = .55$ ($d \approx 0.20$); moderate if $DI_f = .60$ ($d \approx 0.50$); large if $DI_f = .65$ ($d \approx 0.80$); and more than large if $DI_f = .75$ ($d \approx 1.20$). Arce (Fandiño et al., 2021) also proposes the Forensic Error (E_f) for constructing forensic decision models, which estimates the probability of type I error (i.e., false positive). Thus, if the 95% CI of the E_f includes zero, the forensic error is negligible; if the lower limit of the 95% CI of the E_f is $< .05$, the probability of forensic error is trivial (insignificant); if the lower limit of the 95% CI is $\geq .05$, the probability of E_f is significant; if the 95% CI includes or exceeds .50, the E_f is common (i.e., the probability of error is common, $\approx 50\%$); and if the 95% CI includes or exceeds .95, the E_f is normal (i.e., the probability of error is normal, $\geq 95\%$).

Results

Design Sensitivity

Post hoc sensitivity analysis showed that the design sensitivity for the χ^2 test ($\alpha = .05$, $N = 154$, $df = 1$) was .961, that is, the probability of finding a true association between variables is 96.1%. Additionally, the sensitivity of the design for contrasting the observed proportion against a constant (.05) was 93.7%.

Reliability of the Categorical System for Content Analysis of Memories of the GES

The reliability of the GES content category system ($N = 22$; 'improper assumption of responsibility' and 'forgiveness towards

Table 3. Results of the Analysis of the Sufficiency and Validity by Memory Type

Variable	χ^2	p	$f(p)_{liv}$	$f(p)_{fab}$	p_{diff} [95% CI]
Sufficiency	57.75	< .001	77(1.00)	35 (.455)	.545 [.466, .624]
Internal consistency	7.33	.014	77(1.00)	70 (.909)	.091 [.046, .136]
Consistent with prior statement	61.60	< .001	77(1.00)	33 (.429)	.571 [.493, .649]
Temporal and contextual persistence	135.22	< .001	77(1.00)	5 (.065)	.935 [.896, .974]
Consist. scientific and nature laws	1.01	1.00	77(1.00)	76 (.987)	.013 [-.005, .031]
Validity (accumulative effect of the validity categories)	154.00	< .001	77(1.00)	0 (0.00)	1.00 [1.00, 1.00]
Validity and sufficiency	154.00	< .001	77(1.00)	0 (0.00)	1.00 [1.00, 1.00]

Note. $df = 1$; $N = 154$; χ^2 = Fisher's exact test; $f(p)_{liv}$ = frequency (observed proportion) in memories of real-life experiences; $f(p)_{fab}$ = frequency (observed proportion) in fabricated memories of bullying; p_{diff} [95% CI] = probability difference [95% confidence interval].

Table 4. Productivity of the GES Content Analysis Categories in Lived-event Memories

Content category	$f(p_o)$	$Z_{.05}$	$h_{.05}$ [95% CI]	$Z_{.5}$	$h_{.5}$ [95% CI]	$Z_{.95}$	$h_{.95}$ [95% CI]
1. Logical structure and realism	77(1.0)	38.25***	3.80 [3.74, 3.86]	8.77***	2.22 [2.09, 2.35]	2.01*	0.63 [0.58, 0.68]
2. Ellipsis	8(.104)	2.17*	0.29 [0.19, 0.39]	-6.94***	-1.29 [-1.40, -1.18]	-34.06***	-2.88 [-2.96, -2.80]
3. Amount of detail	46(.597)	22.02***	1.86 [1.75, 1.97]	1.70	0.28 [0.21, 0.35]	-14.21***	-1.31 [-1.42, -1.20]
4. Concreteness and clarity of the account	49(.636)	23.59***	1.97 [1.86, 2.08]	2.39*	0.39 [0.31, 0.47]	-12.64***	-1.20 [-1.30, -1.10]
5. Singularity and originality	46(.597)	22.02***	1.86 [1.75, 1.97]	1.70	0.28 [0.21, 0.35]	-14.21***	-1.31 [-1.42, -1.20]
6. Contextual embedding	59(.766)	28.83***	2.37 [2.27, 2.47]	4.66***	0.79 [0.69, 0.89]	-7.41***	-0.80 [-0.89, -0.71]
7. Description of interactions	33(.429)	15.26***	1.38 [1.27, 1.49]	-1.25	-0.20 [-0.26, -0.14]	-20.98***	-1.79 [-1.90, -1.68]
8. Reproduction of conversations	72(.935)	35.63***	3.07 [3.00, 3.14]	7.63***	1.49 [1.38, 1.60]	-0.60	-0.10 [-0.13, -0.07]
9. Narration of unbelievable facts	6(.078)	1.12	0.16 [0.12, 0.20]	-7.41***	-1.42 [-1.53, -1.31]	-35.11***	-3.01 [-3.08, -2.94]
10. Sensory information	44(.571)	20.98***	1.78 [1.67, 1.89]	1.25	0.20 [0.14, 0.26]	-15.26***	-1.39 [-1.50, -1.28]
11. Cognitive operations	32(.416)	14.74***	1.34 [1.23, 1.45]	-1.48	-0.24 [-0.31, -0.18]	-21.50***	-1.83 [-1.94, -1.72]
12. Cognitions inconsistent with the lived experience	5(.065)	0.60	0.09 [0.06, 0.12]	-7.63***	-1.49 [-1.60, -1.38]	-35.63***	-3.08 [-3.15, -3.01]
13. Cognitions about affect and mental state	68(.883)	33.54***	2.82 [2.74, 2.90]	6.72***	1.24 [1.13, 1.35]	-2.70**	-0.35 [-0.41, -0.29]
14. Cognitions about the cause of victimization	13(.169)	4.79***	0.56 [0.49, 0.63]	-5.81***	-1.02 [-1.13, -0.91]	-31.44***	-1.83 [-1.92, -1.74]
15. Cognitions improper assumption responsibility	27(.351)	12.12***	1.15 [1.05, 1.25]	-2.61**	-0.43 [-0.51, -0.35]	-24.12***	-2.02 [-2.13, -1.91]
16. Infrequent information	17(.221)	6.88***	0.74 [0.66, 0.82]	-4.89***	-0.84 [-0.94, -0.74]	-29.35***	-2.43 [-2.53, -2.33]
17. Irrelevant information	41(.532)	19.40***	1.67 [1.56, 1.78]	0.56	-0.42 [-0.46, -0.38]	-16.83***	-1.50 [-1.61, -1.39]
18. Elicitation of memory nodes	17(.221)	6.88***	0.74 [0.66, 0.82]	-4.89***	-0.84 [-0.94, -0.74]	-29.35***	-2.43 [-2.53, -2.33]
19. Spontaneous corrections	18(.234)	7.41***	0.79 [0.71, 0.87]	-4.67***	-0.79 [-0.89, -0.69]	-28.83***	-2.38 [-2.48, -2.28]
20. Admission of memory gaps	25(.325)	11.07***	1.08 [0.98, 1.18]	-3.07**	-0.50 [-0.58, -0.42]	-25.16***	-2.09 [-2.20, -1.98]
21. Expression doubts about one's own testimony	28(.364)	12.64***	1.19 [1.09, 1.29]	-2.39*	-0.39 [-0.47, -0.31]	-23.59***	-1.98 [-2.09, -1.87]
22. Improper assumption of responsibility	11(.143)	3.74***	0.46 [0.40, 0.52]	-6.27***	-1.12 [-1.23, -1.01]	-32.49***	-2.71 [-2.80, -2.62]
23. Forgiveness towards the aggressor	0(.000)	-2.01*	-0.64 [-0.69, -0.59]	-8.77	-2.22 [-2.27, -2.17]	-38.25***	-3.81 [-3.86, -3.76]
24. Subtle clinical symptoms	18(.234)	7.41***	0.79 [0.70, 0.88]	-4.67***	-0.79 [-0.89, -0.69]	-28.83***	-2.38 [-2.48, -2.28]

Note. $f(p)$ = frequency (observed proportion); $Z_{.05}$ = zeta value for the difference between the observed probability and a trivial effect (constant = .05); $Z_{.5}$ = zeta value for the difference between the observed probability and a common effect (constant = .50); $Z_{.95}$ = zeta value for the difference between the observed probability and a general/normal effect (constant = .95); $h_{.05}$ [95% CI] = Cohen's h effect sizes to $Z_{.05}$ [95% confidence interval]; $h_{.5}$ [95% CI] = Cohen's h effect sizes to $Z_{.5}$ [95% confidence interval]; $h_{.95}$ [95% CI] = Cohen's h effect sizes to $Z_{.95}$ [95% confidence interval]; * $p < .05$; ** $p < .01$; *** $p < .001$.

the aggressor' content categories were not assessed due to being unproductive in non-victimization autobiographical memories) in the total sample was excellent, $\alpha = .874$, 95% CI [.843, .901] (the 95% CI of α included .90; George & Mallery, 2024). This indicates that the GES content categories are measuring the same construct (i.e., internal consistency), that is, characteristics of memories of real-life experiences. Given that content categories are indicative of lived experiences, but not of fabricated memories of bullying victimization, reliability analyses were conducted separately for each memory conditions. Results showed a good reliability (the 95% CI of α included .80; George & Mallery, 2024) in memories of lived experiences, $\alpha = .734$, 95% CI [.640, .813], and unacceptable reliability ($\alpha \leq .50$; George & Mallery, 2024) in fabricated memories of bullying victimization, $\alpha = .476$, 95% CI [.289, .632]. That is, the categorical system is not reliable for assessing fabricated memories, with 72.39% of the standard deviation attributable to error; therefore, the GES content analysis categorical system does not allow for the inference of the falsity of the account. These results are consistent with the Undeutsch's (1967) hypothesis and invalidate the use of the system for investigative purposes (e.g., police or judicial investigation), restricting its utility exclusively to testimony evaluation.

Sufficiency and Validity of the Memory (Statement)

The study of the sufficiency of the witness's memory (i.e., sufficient vs. insufficient evidence in judicial and forensic settings) and memory type (lived vs. fabricated) revealed a significant association, $\chi^2(1) = 57.75$, $p < .001$, $\phi = .612$. Sufficient memory was observed in 100% of memories of lived experiences and in 45.5% of fabricated memories of bullying. This corresponds to a probability difference of .545 [.464, .626], which is significant and common (i.e., the 95% CI included .50) and of more than large magnitude, $h = 2.35$, 95% CI [2.24, 2.46].

In turn, validity analyses (see Table 3) reported a significant association between type of memory and memory validity (valid evidence vs. invalid evidence) across all criteria, except for consistency with scientific and natural laws. Specifically, all protocols of memories of real-life experiences met the four validity criteria, whereas fabricated memories of bullying showed inconsistency over time (central information of the account; Montes et al., 2024) in 57.1% of cases and lack of spatio-temporal persistence in 93.5%. Differences in proportions (incremental validity) were significant and more than common ($.50 < p_{\text{diff}} < .95$) for spatio-temporal persistence; significant and common (i.e., 95% confidence interval includes .50) for inter-statement consistency; and significant, but non-common ($.05 < p_{\text{diff}} < .50$) for internal consistency. Accumulatively, all protocols based on memories of lived experiences were both valid and sufficient, whereas none of the fabricated memories jointly met the criteria of sufficiency and validity (general/normal advantage; $p_{\text{diff}} > .95$).

Productivity of the Content Categories

The analyses of the productivity of the categories in memories of real-life experiences (see Table 4) yielded significant results in all categories, except for Narration of unbelievable facts, Cognitions inconsistent with the lived experience and Forgiveness towards the aggressor ($p_0 \leq .05$). Thus, a general/normal prevalence (productivity) ($p_0 \geq .95$) was observed in Logical structure and realism, and in Reproduction of conversations; more than common ($.50 < p_0 < .95$) in Concreteness and clarity of the account, Contextual embedding and Cognitions about affect and mental state; common ($p_0 = .50$) in Amount of detail, Singularity and originality of the account, Description of interactions, Sensory information, Cognitive operations and Irrelevant information; and significant, but less than common ($.05 < p_0 < .50$) in the remaining categories.

Associations between Type of Memory and GES Content Categories

The comparison of the mean number of coded content categories showed significant differences as a function of memory type, $t(101.9) = 16.22$, $p < .001$, such that autobiographical memories contain more content categories ($M = 9.87$) than fabricated memories of bullying victimization ($M = 2.58$). The effect size was more than large, $d = 2.60$, and the probability of error of the statistical model (i.e., that a memory of lived experiences contains fewer categories than the mean of fabricated memories of bullying victimization) was 0.4% (PIS = .004).

Concerning the "general characteristics" (see Table 5), all categories showed a significant association with memory type, with a more than large effect size ($PR > 8.82$) for Singularity and originality of the account, Amount of detail, and Concreteness and clarity of the account; a large effect size ($4.25 < PR < 8.82$) for Ellipsis; and a negligible effect size ($PR < 1.44$) for Logical structure and realism. In sum, the Forensic Discrimination Index (DI_f) showed a more than large effect size ($DI_f > .75$) for Amount of detail, Concreteness and clarity of the account, and Singularity and originality of the account, whereas the estimation of the Forensic Error (E_f ; probability of false positive, punishable error in the forensic context) indicated a negligible probability of error (95% CI includes $E_f = 0$) in all categories except Logical structure and realism, for which the error was normal ($E_f = .95$; normal error).

In the "memory contents" (see Table 5), a significant association was observed between memory type and all content categories, except Narration of unbelievable facts (non-productive category in memories of lived experiences). The effect size was more than large ($PR > 8.82$) for Contextual, spatial, and temporal embedding and Sensory information; large ($4.25 < PR < 8.82$) for Description of interactions; and small ($1.44 < PR < 2.47$) for Reproduction of conversations. The discrimination capacity between memories was more than large ($DI_f > .75$) for Contextual, spatial, and temporal embedding and Sensory information; and large ($.65 < DI_f < .75$) for Description of interactions and Reproduction of conversations. The probability of forensic error was negligible (95% CI includes $E_f = .05$) in all categories except Reproduction of conversations, for which it was common (95% CI includes $E_f = .50$; common error).

In the "witness cognitions" (see Table 5), the results revealed a significant association between the content categories and memories of real-life experiences, except Cognitions inconsistent with the lived experience (non-productive category in memories of lived experiences). The effect size was more than large ($PR > 8.82$) for Cognitive operations, Cognitions about the cause of victimization, and Cognitions of improper assumption of responsibility, and small ($1.44 < PR < 2.47$) for Cognitions about affect and mental state. The discrimination capacity between memories was large ($.65 < DI_f < .75$) for Cognitive operations, Cognitions about affect and mental state, and Cognitions of improper assumption of responsibility; and small ($.55 < DI_f < .60$) for Cognitions about the cause of victimization. The probability of forensic error was negligible ($E_f = 0$) in all categories except in Cognitions about affect and mental state, for which it was common (95% CI of the E_f included .50; common error).

Regarding the "content characteristics" (see Table 5), the results showed a significant association between the content categories and memories of lived experiences with a more than large magnitude ($PR > 8.82$) for Irrelevant information and Elicitation of memory nodes; and large ($4.25 < PR < 8.82$) for Infrequent information. The discrimination capacity between memories was more than large ($DI_f > .75$) for Irrelevant information; moderate ($.60 < DI_f < .65$) for Elicitation of memory nodes; and small ($.55 < DI_f < .60$) for Infrequent information. The probability of forensic error was negligible in all categories (95% CI includes $E_f = 0$).

With respect to the "memory quality contents" (see Table 5), the results did not support a significant association between memories of real-life experiences and Admission of memory gaps and Forgiveness

Table 5. Associations between Memory Type and GES Content Analysis Categories

Content category	χ^2	p	PR [95% CI]	$f(p)_{liv}$	$f(p)_{fab}$	DI _f [95% CI] E _f [95% CI]
General characteristics						
1. Logical structure and realism	9.56	.003	1.13 [1.04, 1.23]	77 (.10)	68 (.883)	.558 [.480, .636] .883 [.811, .954]
2. Ellipsis	5.78	.034	8.00 [1.02, 62.44]	8 (.104)	1 (.013)	.545 [.466, .624] .013 [-.012, .038]
3. Amount of detail	65.59	< .001	46.00 [-----]	46 (.597)	0 (.000)	.799 [.736, .862] .000 [.000, .000]
4. Concreteness and clarity of the account	71.87	< .001	49.00 [-----]	49 (.636)	0 (.000)	.818 [.757, .879] .000 [.000, .000]
5. Singularity and originality	55.34	< .001	15.33 [4.98, 47.20]	46 (.597)	3 (.039)	.779 [.713, .845] .039 [-.004, .082]
Memory contents						
6. Contextual embedding	91.85	< .001	59.00 [8.38, 415.16]	59 (.766)	1 (.013)	.876 [.824, .928] .013 [-.012, .038]
7. Description of interactions	27.39	< .001	6.60 [2.72, 16.01]	33 (.429)	5 (.065)	.682 [.608, .756] .065 [.001, .120]
8. Reproduction of conversations	31.94	< .001	1.76 [1.41, 2.18]	72 (.935)	41 (.532)	.701 [.629, .773] .532 [.421, .653]
9. Narration of unbelievable facts	0.84	.772	0.86 [0.30, 2.43]	6 (.078)	7 (.091)	.494 [.415, .573] .091 [.027, .155]
10. Sensory information	54.68	< .001	22.00 [5.53, 87.57]	44 (.571)	2 (.026)	.773 [.707, .839] .026 [-.001, .062]
Witness cognitions						
11. Cognitive operations	40.39	< .001	32.00 [-----]	32 (.416)	0 (.000)	.708 [.636, .780] .000 [.000, .000]
12. Cognitions inconsistent with the lived experience	5.17	.058	5.00 [-----]	5 (.065)	0 (.000)	.532 [.453, .611] .000 [.000, .000]
13. Cognitions about affect and mental state	30.33	< .001	1.89 [1.47, 2.43]	68 (.883)	36 (.468)	.708 [.636, .780] .468 [.357, .579]
14. Cognitions about the cause of victimization	14.20	< .001	13.00 [-----]	13 (.169)	0 (.000)	.584 [.506, .662] .000 [.000, .000]
15. Cognitions improper assumption responsibility	32.74	< .001	27.00 [-----]	27 (.351)	0 (.000)	.675 [.601, .749] .000 [.000, .000]
Content characteristics						
16. Infrequent information	11.26	< .001	5.67 [1.73, 18.55]	17 (.221)	3 (.039)	.591 [.513, .669] .039 [-.004, .082]
17. Irrelevant information	52.38	< .001	41.00 [5.78, 290.62]	41 (.532)	1 (.013)	.760 [.693, .827] .013 [-.012, .038]
18. Elicitation of memory nodes	16.11	< .001	17.00 [2.32, 124.61]	17 (.221)	1 (.013)	.604 [.527, .681] .013 [-.012, .038]
Content regarding memory quality						
19. Spontaneous corrections	17.35	< .001	18.00 [2.46, 131.52]	18 (.234)	1 (.013)	.610 [.533, .687] .013 [-.012, .03]
20. Admission of memory gaps	3.38	.066	1.67 [0.95, 2.91]	25 (.325)	15 (.195)	.565 [.487, .643] .195 [.107, .283]
21. Expression doubts about one's own testimony	27.99	< .001	14.00 [3.45, 56.74]	28 (.364)	2 (.026)	.669 [.595, .743] .026 [-.010, .062]
22. Improper assumption of responsibility	6.81	.009	5.50 [1.26, 23.99]	11 (.143)	2 (.026)	.558 [.480, .636] .026 [-.010, .062]
23. Forgiveness towards the aggressor	2.03	.497	5.00 [-----]	0 (.000)	2 (.026)	.487 [.400, .590] .026 [-.010, .062]
Psychological harm derived from the action under judgement						
24. Subtle clinical symptoms	4.63	.031	2.25 [1.04, 4.86]	18 (.234)	8 (.104)	.565 [.487, .643] .104 [.036, .172]

Note. $df = 1$, $N = 154$; χ^2 = Fisher's exact test; ϕ = phi correlation, effect size; PR [95% CI] = prevalence ratio [95% confidence interval]; $f(p)_{liv}$ = frequency (observed proportion) in memories of real-life experiences; $f(p)_{fab}$ = frequency (observed proportion) in bullying fabricated memories; DI_f [95% CI] = forensic discrimination index [95% confidence interval]; E_f [95% CI] = forensic error [95% confidence interval]. When a content category was absent in the fabricated memory condition, the PR corresponds to the prevalence in the autobiographical memory condition. The 95% CI is not estimable because the denominator is zero.

towards the aggressor (not coded in memories of lived experiences). Conversely, the association was significant and with a more than large effect size ($PR > 8.82$) for Spontaneous corrections, specifications and elaborations of the account, and Expression of doubts about one's own testimony; and large ($4.25 < PR < 8.82$) for Improper assumption of responsibility. Additionally, the discrimination capacity between memories was large ($.65 < DI_f < .75$) for Expression of doubts about

one's own testimony; and moderate ($.60 < DI_f < .65$) for Spontaneous corrections, specifications and elaborations of the account. Moreover, in these categories the observed forensic error was negligible (95% CI includes $E_f = 0$).

Finally, it was found that Subtle clinical symptoms are significantly associated with memory type (see Table 5), with a small effect size ($1.44 < PR < 2.47$), a small discrimination capacity be-

Table 6. Forensic Decision Models

Complete model	Higher prevalence model	Robust model
1. Logical structure and realism	1. Logical structure and realism	2. Ellipsis
2. Ellipsis	2. Ellipsis	3. Amount of detail
3. Amount of detail	3. Amount of detail	4. Concreteness and clarity of the account
4. Concreteness and clarity of the account	4. Concreteness and clarity of the account	5. Singularity and originality
5. Singularity and originality	5. Singularity and originality	6. Contextual embedding
6. Contextual embedding	6. Contextual embedding	7. Description of interactions
7. Description of interactions	7. Description of interactions	10. Sensory information
8. Reproduction of conversations	8. Reproduction of conversations	11. Cognitive operations
9. Narration of unbelievable facts	10. Sensory information	14. Cognitions about the cause of victimization
10. Sensory information	11. Cognitive operations	15. Cognitions improper assumption responsibility
11. Cognitive operations	13. Cognitions about affect and mental state	16. Infrequent information
12. Cognitions inconsistent with the lived experience	14. Cognitions about the cause of victimization	17. Irrelevant information
13. Cognitions about affect and mental state	15. Cognitions improper assumption responsibility	18. Elicitation of memory nodes
14. Cognitions about the cause of victimization	16. Infrequent information	19. Spontaneous corrections
15. Cognitions improper assumption responsibility	17. Irrelevant information	21. Expression doubts about one's own testimony
16. Infrequent information	18. Elicitation of memory nodes	22. Improper assumption of responsibility
17. Irrelevant information	19. Spontaneous corrections	24. Subtle clinical symptoms
18. Elicitation of memory nodes	21. Expression doubts about one's own testimony	
19. Spontaneous corrections	22. Improper assumption of responsibility	
20. Admission of memory gaps	24. Subtle clinical symptoms	
21. Expression doubts about one's own testimony		
22. Improper assumption of responsibility		
23. Forgiveness towards the aggressor		
24. Subtle clinical symptoms		

tween memories ($.55 < DI_f < .60$), and a trivial probability of forensic error (the 95% CI for E_f included .05).

Forensic Decision Criteria

Although the GES categorical system has demonstrated good reliability, and its content categories are generally significantly associated with memories of real-life experiences, the associated error is bidirectional, affecting both false positives (i.e., assigning probative value to fabricated testimony) and false negatives (i.e., failing to assign probative value to honest accounts). However, forensic evaluation (assistance role to justice) in the justice system requires that type I error (false positives) be zero, as it violates the principle of presumption of innocence (i.e., no innocent person should be convicted; *Sentencia 213/2002* del Tribunal Supremo [sentence of the Spanish Supreme Court]). Therefore, scientific findings are not directly generalisable to forensic practice, and strict decision criteria (criterial validity) must be defined based on scientific evidence to control the subjectivity of the forensic expert based on clinical judgement or on decision rules without scientific support (Köhnken, 2004; Steller, 1989; Undeutsch, 1989). This ensures strict adherence to the principle of presumption of innocence, given the “nuclear evidence” nature of testimony evaluation (i.e., evidence on which the burden of proof rests and without which the remaining evidence lacks sufficient probative value; Arce et al., 2015). To this end, the GES includes a procedure that verifies the maximum number of content categories in fabricated memories of bullying victimization, which serves as a strict decision criterion (criterial validity) for compliance with the principle of presumption of innocence, across three decision models (see Table 6): a) Complete model, with the 24 original GES categories; b) Higher prevalence model, with 20 categories, comprising those categories significantly associated with memories of real-life experiences; and c) Robust model, composed of 17 categories that are associated with a robust effect size ($PR \geq 2.47$) in memories of lived experiences.

Using the Complete model (see Table 7), a maximum of 9 content categories were coded in fabricated memories of bullying victimization; therefore, the registration of 10 or more content categories (criterial validity) supports classification as a memory of lived experience, ensuring strict compliance with the principle of presumption of innocence. In the Higher prevalence model, a maximum of 7 categories was observed; thus, the criterial validity

is the registration of ≥ 8 , and, following the same procedure, ≥ 5 for the Robust model. Comparatively, there were no differences in performance between the Complete and higher Prevalence models, $Z(N = 44) = -0.17, p = .865$; the Complete and Robust models $Z(N = 41) = -0.85, p = .395$; and the Higher prevalence and Robust models, $Z(N = 37) = -0.65, p = .516$.

Table 7. Decision Criteria in Forensic Assessment

Number of Categories	Complete model	Higher-prevalence model	Robust model
0	1 (1.3%)	1 (1.3%)	57 (74.0%)
1	19 (24.7%)	21 (28.6%)	14 (92.2%)
2	19 (50.6%)	28 (64.9%)	3 (96.1%)
3	24 (81.8%)	19 (89.6%)	2 (98.7%)
4	7 (90.9%)	3 (93.5%)	1 (100%)
5	3 (94.8%)	2 (96.1%)	
6	3 (98.7%)	2 (98.7%)	
7		1 (100%)	
8			
9	1 (100%)		

Discussion

The fidelity of the coding process, defined as adherence of category coding to the manual, and its objectivity, defined as the control of rater subjectivity in coding, are supported by the very good between- and within-rater concordance and agreement, thereby validating the coding manual in line with the Daubert standard.

Overall, the findings support the association between the GES content categories and memories of real-life experiences (Arce, 2024; Arce & Fariña, 2005, 2006, 2014, 2015). However, the results for Narration of unbelievable facts, Cognitions about the cause of victimization, and Forgiveness towards the aggressor should not be generalized, as these categories were either not coded or had a low probability of occurrence in the non-criminal autobiographical memories used in this study. That is, they validate the GES content analysis categorical system for the classification of memories of real-life experiences. The effects for the categories Narration of unbelievable facts, Cognitions about the cause of victimization and Forgiveness towards the aggressor are not generalizable in this

study because they are not registrable or have a low probability in memories of real-life experiences that are not victimization of a crime. Conversely, no association was observed between GES content categories and fabricated memories of bullying victimization; proving that the GES content categories are not valid for classifying fabricated memories. The system is not reliable with fabricated memories of bullying victimization. Summing up, the GES content analysis categorical system is neither reliable nor valid for the classification of fabricated memories of bullying victimization and, therefore, cannot be used to test alternative hypotheses to memories of lived experiences. Consequently, it is not suitable for investigative tasks (e.g., judicial or police investigations), and its evidential value in judicial settings is limited to providing probative value to the complainant's testimony. These results also sustain the validity of the Undeutsch hypothesis on which the GES is based: memories of truthful accounts differ qualitatively from fabricated ones, and the presence of reality criteria is associated with truthful memories, whereas their absence cannot be related to fabricated memories. These findings are consistent with previous research demonstrating the validity and reliability of the GES content analysis system in other victimization contexts, such as intimate partner violence (Arce et al., 2013; Pérez, 2024; Vilariño et al., 2011), sexual assaults against adults (Arce, 2024), child sexual abuse (Abeledo, 2024) or workplace bullying (Arce & Fariña, 2011). These studies, combining high-fidelity laboratory and field research designs, involve all types of populations and case types. Consequently, the available evidence supports the potential generalizability of the GES content category system across different populations and case types.

The results established strict decision criteria (evidence-based criterial validity) that guarantees compliance with the principle of presumption of innocence (type I error/false positives = 0): registration of ≥ 10 content categories in the complete model, ≥ 8 in the higher-prevalence model, and ≥ 5 in the robust model. Given that the models did not differ in performance, any of these criteria may be used (or, in any case, the one that is met). Future research with actual field cases of bullying victimization should establish the type II error (i.e., false negative) associated with each decision model.

The GES also assesses sufficiency (memory) and evidence validity (consistency) of the testimony. Unlike in previous studies, all fabricated memories of bullying victimization were coded as insufficient and/or invalid evidence. However, this finding should not be interpreted as evidence of false memory or testimony. According to the GES forensic technique, derived from judicial decisions and case law (i.e., construct validity), invalid (suggested and implanted memories are invalid; Arce et al., 2023; Villarquide et al., 2025) or insufficient evidence (an insufficient or invalid memory may contain content criteria to reach criterial validity potentially overlapping with fabricated accounts) does not support the complainant's testimony.

From an applied perspective, these findings have direct implications for the forensic evaluation of bullying. When content analysis supports that the statement is based on a memory of a lived experience and the evidence (i.e., testimony) is both sufficient and valid, the reported events may assist in determining whether the experience fulfils the diagnostic criteria for bullying (Olweus & Limber, 2018; P. K. Smith et al., 2012) or corresponds to another form of victimization, such as dating violence, sexual violence, child maltreatment, physical fights, rough play, or hurtful teasing. Establishing whether the reported events correspond to bullying is essential for selecting appropriate psychosocial interventions, implementing school-based protection measures, and planning evidence-based support for both victims and perpetrators (Gaffney et al., 2019, 2021; Hensums et al., 2023). Although psychometric instruments are valuable for screening and diagnostic purposes, they should be complemented by behavioral observations and interviews about the alleged events within a multimethod assessment framework (Arce et al., 2014; Freeman et al., 2012). Furthermore, when cases require legal or administrative

action, scientifically supported testimony evaluation may contribute to decision-making while reducing the risk of false-positive conclusions and safeguarding the presumption of innocence.

The findings of this study are subject to limitations that should be had in mind when generalizing the results. First, to ensure ethical safeguards and prevent revictimization, accounts of bullying victimization were not collected in the autobiographical memory condition; instead, participants reported negative events involving a loss of control over the situation and its outcomes. Consequently, memories of lived experiences did not correspond to genuine bullying experiences, and the study followed a high-fidelity simulation design (Rogers, 2018), in which autobiographical memories served as the control condition whereas fabricated allegations of bullying constituted the experimental condition. Although such designs offer strong internal validity, Consequently, memories of real-life experiences do not correspond to lived events of bullying (high-fidelity simulation design). Although such designs offer strong internal validity, they may be limited in external validity (Rogers, 2018), and part of the observed effect may be attributable to the method rather than the construct (Podsakoff et al., 2003). Second, the GES content categories of memories of lived experiences (Narration of unbelievable facts, Cognitions about the cause of victimization, and Forgiveness towards the aggressor) showed reduced productivity in memories of real-life experiences due to the nature of the experimental condition, that is, a directly experienced negative event involving loss of control over the situation and outcomes. Likewise, the GES categories specific to bullying cases, minimization of bullying or its consequences and attribution of blame to the victim, were not analyzed because the autobiographical memory was not of bullying. This resulted in lower observed reliability of the system and, consequently, reduced validity.

Conflict of Interest

The authors of this article declare no conflict of interest.

References

- Abeledo, M. (2024). *Estudio de la eficacia discriminativa y de clasificación entre la memoria de hechos auto-experimentados y fabricados de los criterios de contenido de la declaración del SEG en casos de violencia sexual infantil y adolescente* [Doctoral dissertation, University of Santiago de Compostela]. <https://minerva.usc.gal/bitstreams/209e863f-5d7d-41e3-a16a-a327d49d0d6c/download>
- Ak, S., Özdemir, Y., & Kuzucu, Y. (2015). Cybervictimization and cyberbullying: The mediating role of anger, don't anger me! *Computers in Human Behavior*, 49, 437–443. <https://doi.org/10.1016/j.chb.2015.03.030>
- Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *European Journal of Psychology Applied to Legal Context*, 7(1), 3–12. <https://doi.org/10.1016/j.ejpal.2014.11.002>
- American Psychiatric Association (APA, 2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text revised). American Psychiatric Association. <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 forense [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.14.2.21347>
- Arce, R. (2024). Carga de la prueba con el Sistema de Evaluación Global ajustado a casos recogidos en la Ley Orgánica 10/2022, de 6 de septiembre, de garantía integral de la libertad sexual. In D. Seijo, D. Pineda, & F. Fariña, F. (Eds.), *Nuevas aportaciones basadas en evidencia científica a la práctica de la psicología jurídica y forense. Colección Psicología y Ley*, 18, 221–236. Universidad Miguel Hernández de Elche.
- 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. <http://www.papelesdelpsicologo.es/pdf/1247.pdf>

- Arce, R., & Fariña, F. (2006). Psicología del testimonio: Evaluación de la credibilidad y de la huella psíquica en el contexto penal. In Consejo General del Poder Judicial (Ed.), *Psicología del testimonio y prueba pericial* (pp. 39-103). Consejo General del Poder Judicial.
- Arce, R., & Fariña, F. (2011). Evaluación forense del acoso moral en el trabajo (Mobbing) mediante el Sistema de Evaluación Global. In M. Matos, R. Abrunhosa, & C. Machado (Eds.), *Manual de psicología forense: contextos, prácticas e desafíos* (pp. 375-398). Psiquilibrios Edicoes.
- Arce, R., & Fariña, F. (2014). Práctica de la prueba psicológica-forense: El Sistema de Evaluación Global (SEG). In C. Chan, C. Estrada, & F. J. Rodríguez (Eds.), *Aportaciones a la psicología jurídica y forense desde Iberoamérica* (pp. 47-61). Manual Moderno. <https://www.researchgate.net/publication/274389353>
- Arce, R., & Fariña, F. (2015). Evaluación psicológico-forense de la credibilidad y daño psíquico mediante el Sistema de Evaluación Global. In P. Rivas & G. L. Barrios (Eds.), *Violencia de género: perspectiva multidisciplinar y práctica forense* (pp. 411-441). Thomson Aranzadi.
- Arce, R., Fariña, F., & Fraga, A. (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.psicothema.com/pdf/381.pdf>
- Arce, R., Fariña, F., Novo, M., & Vilariño, M. (2013). Contraste de la efectividad en la detección de declaraciones fabricadas de los criterios de validez del Sistema de Evaluación Global en casos de violencia contra la mujer. In F. Expósito, I. Valor-Segura, M. Vilariño, & A. Palmer (Eds.), *Psicología jurídica aplicada a los problemas sociales. Colección Psicología y Ley*, 11, 151-159. Sociedad Española de Psicología Jurídica y Forense. <https://doi.org/10.13140/RG.2.1.3957.1363>
- Arce, R., Fariña, F., & Vilariño, M. (2015). Daño psicológico en casos de víctimas de violencia de género: 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. <http://dx.doi.org/10.1016/j.rips.2015.04.002>
- 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(4), 100386. <https://doi.org/10.1016/j.ijchp.2023.100386>
- Arce, R., Velasco, J., Novo, M., & Fariña, F. (2014). Elaboración y validación de una escala para la evaluación del acoso escolar [Development and validation of a scale for the evaluation of school bullying]. *Revista Iberoamericana de Psicología y Salud*, 5(1), 71-104. <https://doi.org/10.70478/rips.2014.05.05>
- Bradshaw, C. P., Waasdorp, T. E., & Johnson, S. L. (2015). Overlapping verbal, relational, physical, and electronic forms of bullying in adolescence: Influence of school context. *Journal of Clinical Child & Adolescent Psychology*, 44(3), 494-508. <https://doi.org/10.1080/15374416.2014.893516>
- Castro-Gallego, M. P., Carvajal-Builes, J. C., & Jiménez-Molina, J. R. (2024). Sistema de evaluación global: una revisión narrativa. In J. C. Carvajal-Builes, J. R. Jiménez-Molina, E. Devia-Vega, E. J. López-Cantero, I. A. Rodríguez-Montealegre, L. T. Muñoz-Castro, Y. S. Chaparro-Gómez, & M. P. Castro-Gallego, *Análisis del comportamiento verbal: Estrategias en psicología del testimonio* (pp. 39-54). Editorial Universidad Católica de Colombia. <https://doi.org/10.14718/9786287554733.2024.3>
- Cea, B., Montes, A., Trinidad, A., & Arce, R. (2025). Forensic psychological harm in victims of child sexual abuse: A meta-analytic review. *European Journal of Psychology Applied to Legal Context*, 7(2), 111-129. <https://doi.org/10.5093/ejpalc2025a10>
- Ceci, S. J., Kulkofsky, S., Klemfuss, J. Z., Sweeney, C. D., & Bruck, M. (2007). Unwarranted assumptions about children's testimonial accuracy. *Annual Review of Clinical Psychology*, 3, 311-328. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091354>
- Centers for Disease Control and Prevention (CDC, 2022). *Preventing teen dating violence: What is teen dating violence?* https://stacks.cdc.gov/view/cdc/129418/cdc_129418_DS1.pdf&ved=2ahUKEwiOgZfM-baNAxVf_srlHezkj_gQFnoECA5QAQ&usg=AOvVaw3_tFhJL9Lk1bI0N9VjkzV9
- Chun, J., Lee, J., Kim, J., & Lee, S. (2020). An international systematic review of cyberbullying measurements. *Computers in Human Behavior*, 113, Article 106485. <https://doi.org/10.1016/j.chb.2020.106485>
- Cohen, J. B. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). LEA.
- Colwell, K., Hiscock, C. K., & Memon, A. (2002). Interviewing techniques and the assessment of statement credibility. *Applied Cognitive Psychology*, 16(3), 287-300. <https://doi.org/10.1002/acp.788>
- Cosma, A., Walsh, S. D., Chester, K. L., Callaghan, M., Molcho, M., Craig, W., & Pickett, W. (2020). Bullying victimization: Time trends and the overlap between traditional and cyberbullying across countries in Europe and North America. *International Journal of Public Health*, 65(1), 75-85. <https://doi.org/10.1007/s00038-019-01320-2>
- Danzi, B. A., & La Greca, A. M. (2016). DSM-IV, DSM-5, and ICD-11: Identifying children with posttraumatic stress disorder after disasters. *Journal of Child Psychology and Psychiatry*, 57(12), 1444-1452. <https://doi.org/10.1111/jcpp.12631>
- Daubert vs. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).
- Davis, D., & Loftus, E. F. (2007). Internal and external sources of misinformation in adult witness memory. In M. P. Toglia, J. D. Read, D. F. Ross, & R. C. Lindsay (Eds.), *The handbook of eyewitness psychology: Volume 1: Memory for events* (pp. 195-237). Routledge.
- Davis, J. P., Merrin, G. J., Ingram, K. M., Espelage, D. L., Valido, A., & El Sheikh, A. J. (2019). Examining pathways between bully victimization, depression, and school belonging among early adolescents. *Journal of Child and Family Studies*, 28, 2365-2378. <https://doi.org/10.1007/s10826-019-01340-9>
- Davis, M. R., McMahon, M., & Greenwood, K. M. (2005). The efficacy of mnemonic components of the cognitive interview: Towards a shortened variant for timecritical investigations. *Applied Cognitive Psychology*, 19(1), 75-93. <https://doi.org/10.1002/acp.1048>
- Deeb, H., Vrij, A., Leal, S., Giorgianni, D., Hypšová, P., & Mann, S. (2024). Exposing suspects to their sketches in repeated interviews to elicit information and veracity cues. *European Journal of Psychology Applied to Legal Context*, 16(1), 1-15. <https://doi.org/10.5093/ejpalc2024a1>
- Díaz-Aguado, M. J., Martínez-Arias, R., Falcón, L., & Alvarino, M. (2024). *Acoso escolar y ciberacoso en España en la infancia y en la adolescencia*. Editorial Pirámide.
- Dodier, O., & Otgaar, H. (2019). The forensic and clinical relevance of evidence-based investigative interview methods in historical sexual abuse cases. *Clinical Psychological Science*, 7(6), 1244-1248. <https://doi.org/10.1177/2167702619858287>
- Englander, E., Donnerstein, E., Kowalski, R., Lin, C. A., & Parti, K. (2017). Defining cyberbullying. *Pediatrics*, 140(2), S148-S151. <https://doi.org/10.1542/peds.2016-1758U>
- 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/fpsyg.2021.616855>
- Fariña, F., Arce, R., & Novo, M. (2002). Heurístico de anclaje en las decisiones judiciales [Anchorage in judicial decision making]. *Psicothema*, 14(1), 39-46. <https://www.psicothema.com/pdf/684.pdf>
- Fariña, F., Arce, R., & Real, S. (1994). Ruedas de identificación: De la simulación y la realidad [Lineups: A comparison of high fidelity research and research in a real context]. *Psicothema*, 7(1), 395-402. <http://www.psicothema.es/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, Article e32. <https://doi.org/10.1017/sjp.2014.30>
- Fisher, R. P., & Geiselman, R. E. (1992). *Memory enhancing techniques for investigative interviewing: The cognitive interview*. Charles C. Thomas.
- Fisher, R. P., & Geiselman, R. E. (2010). The cognitive interview method of conducting police interviews: Eliciting extensive information and promoting therapeutic jurisprudence. *International Journal of Law and Psychiatry*, 33(5-6), 321-328. <https://doi.org/10.1016/j.ijlps.2010.09.004>
- Fisher, R. P., Geiselman, R. E., & Amador, M. (1989). Field test of the cognitive interview: Enhancing the recollection of actual victims and witness of crime. *Journal of Applied Psychology*, 74(5), 722-727. <https://doi.org/10.1037/0021-9010.74.5.722>
- Freeman, B., Thompson, C., & Jaques, C. (2012). Forensic aspects and assessment of school bullying. *Psychiatric Clinics of North America*, 35(4), 877-900. <https://doi.org/10.1016/j.psc.2012.08.007>
- Gaffney, H., Tofli, M. M., & Farrington, D. P. (2019). Evaluating the effectiveness of school bullying prevention programs: An updated meta-analytical review. *Aggression and Violent Behavior*, 45, 111-133. <https://doi.org/10.1016/j.avb.2018.07.001>
- Gaffney, H., Tofli, M. M., & Farrington, D. P. (2021). What works in anti-bullying programs? Analysis of effective intervention components. *Journal of School Psychology*, 85, 37-56. <https://doi.org/10.1016/j.jsp.2020.12.002>
- Garaigordobil, M., & Machimbarrena, J. M., (2019). Victimization and perpetration of bullying/cyberbullying: Connections with emotional and behavioral problems and childhood stress. *Psychosocial Intervention*, 28(2), 67-73. <https://doi.org/10.5093/pi2019a3>
- George, D., & Mallery, P. (2024). *IBM SPSS statistics 29 step by step: A simple guide and reference*. Routledge.
- Giorgianni, D., Vrij, A., Leal, S., & Deeb, H. (2025). The effect of the interviewer's cognitive load on the quality of the forensic interview. *European Journal of Psychology Applied to Legal Context*, 17(2), 101-110. <https://doi.org/10.5093/ejpalc2025a9>
- Hensmans, M., De Mooij, B., Kuijper, S. C., Fekkes, M., & Overbeek, G. (2023). What works for whom in school-based anti-bullying interventions? An individual participant data meta-analysis. *Prevention Science*, 24(8), 1435-1446. <https://doi.org/10.1007/s11121-022-01387-z>
- Hong, J. S., Espelage, D. L., & Lee, J. M. (2018). School climate and bullying prevention programs. In H. Saphiro (Ed.), *The Wiley handbook on violence in education: Forms, factors, and preventions* (pp. 359-374). Wiley. <https://doi.org/10.1002/9781118966709.ch17>
- Idsoe, T., Vaillancourt, T., Dyregrov, A., Hagen, K. A., Ogden, T., & Nærde, A. (2021). Bullying victimization and trauma. *Frontiers in Psychiatry*, 11, Article 480353. <https://doi.org/10.3389/fpsyg.2020.480353>

- Katsaras, G. N., Vouloumanou, E. K., Kourlaba, G., Kyritsi, E., Evagelou, E., & Bakoula, C. (2018). Bullying and suicidality in children and adolescents without predisposing factors: A systematic review and meta-analysis. *Adolescent Research Review*, 3, 193-217. <https://doi.org/10.1007/s40894-018-0081-8>
- Kim, J., Song, H., & Jennings, W. G. (2017). A distinct form of deviance or a variation of bullying? Examining the developmental pathways and motives of cyberbullying compared with traditional bullying in South Korea. *Crime & Delinquency*, 63(12), 1600-1625. <https://doi.org/10.1177/0011128716675358>
- Köhnken, G. (2004). Statement validity analysis and the detection of the truth. In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 41-63). Cambridge University Press. <https://doi.org/10.1017/CBO9780511490071.003>
- Konecni, V., & Ebbesen, E. (1992). Methodological issues on legal decision making, with special reference to experimental simulations. In F. Lösel, D. Bender, & T. Bliesener (Eds.), *Psychology and law. International perspectives* (pp. 413-423). Walter de Gruyter.
- Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, 140(4), Article 1073. <https://doi.org/10.1037/a0035618>
- Kwan, I., Dickson, K., Richardson, M., MacDowall, W., Burchett, H., Stansfield, C., Brunton, G., Sutcliffe, K., & Thomas, J. (2020). Cyberbullying and children and young people's mental health: A systematic map of systematic reviews. *Cyberpsychology, Behavior, and Social Networking*, 23(2), 72-82. <https://doi.org/10.1089/cyber.2019.0370>
- Laith, R., & Vaillancourt, T. (2022). The temporal sequence of bullying victimization, academic achievement, and school attendance: A review of the literature. *Aggression and Violent Behavior*, 64, Article 101722. <https://doi.org/10.1016/j.avb.2022.101722>
- Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. <http://doi.org/10.2307/2529310>
- Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de derechos digitales. (2018). *Boletín Oficial del Estado*, 294, 119788-119857. <https://www.boe.es/boe/dias/2018/12/06/pdfs/BOE-A-2018-16673.pdf>
- Li, C., Wang, P., Martin-Moratinos, M., Bella-Fernández, M., & Blasco-Fontecilla, H. (2024). Traditional bullying and cyberbullying in the digital age and its associated mental health problems in children and adolescents: A meta-analysis. *European Child & Adolescent Psychiatry*, 33(9), 2895-2909. <https://doi.org/10.1007/s00787-022-02128-x>
- Lucas-Molina, B., Pérez-Albéniz, A., Solbes-Canales, I., Ortuño-Sierra, J., & Fonseca-Pedrero, E. (2022). Bullying, cyberbullying and mental health: The role of student connectedness as a school protective factor. *Psychosocial Intervention*, 31(1), 33-41. <https://doi.org/10.5093/pi2022a1>
- Marcos, V., Montes, Á., Cea, B., & Seijo, M. D. (2024). Prevalence and quantifying the effects of bullying victimization in school-aged children in internalizing and externalizing mental health problems and academic failure. *Revista de Psicología Clínica con Niños y Adolescentes*, 11(2), Article 8. <https://doi.org/10.21134/rpcna.2024.11.2.8>
- Marcos, V., Seijo, D., Montes, Á., & Arce, R. (2024). Prevalence and quantification of the effects of sexual harassment victimization of school-aged adolescents. *Children*, 11(1), Article 23. <https://doi.org/10.3390/children11010023>
- Martínez, M. S. (2024). *Evaluación de la credibilidad y huella psicológica en casos de violencia de género: Adaptación del Sistema de Evaluación Global al contexto colombiano* [Doctoral dissertation, University of Santiago de Compostela]. <https://minerva.usc.gal/bitstreams/5581de8b-04ce-4904-8931-6bb0ba597c9c/download>
- Martínez, N., & Orihuela, M. J. (2020). Aplicación del sistema de evaluación global en la valoración de la credibilidad en adultos. *Almoraima. Revista de Estudios Campogibraltareños*, 44, 101-107.
- Memon, A., Meissner, C. A., & Fraser, J. (2010). Cognitive interview. A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, and Law*, 16(4), 340-372. <https://doi.org/10.1037/a0020518>
- Mena, C. G. (2021). El Sistema de Evaluación Global (SEG) como instrumento de valoración de la veracidad del testimonio [The global evaluation system (GES) as an instrument for assessing the veracity of the testimony]. *South Florida Journal of Development*, 2(3), 4505-4515. <https://doi.org/10.46932/sfjdv2n3-053>
- Modecki, K. L., Minchin, J., Harbaugh, A. G., Guerra, N. G., & Runions, K. C. (2014). Bullying prevalence across contexts: A meta-analysis measuring cyber and traditional bullying. *Journal of Adolescent Health*, 55(5), 602-611. <https://doi.org/10.1016/j.jadohealth.2014.06.007>
- Montes, A., Sanmarco, J., Cea, B., Marcos, V., & Seijo, D. (2024). Examining the persistence of central and peripheral information in honest witness. *Revista Iberoamericana de Psicología y Salud*, 15(2), 96-104. <https://doi.org/10.23923/j.rips.2024.02.079>
- 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(21), Article 13852. <https://doi.org/10.3390/ijerph192113852>
- Moore, S. E., Norman, R. E., Suetani, S., Thomas, H. J., Sly, P. D., & Scott, J. G. (2017). Consequences of bullying victimization in childhood and adolescence: A systematic review and meta-analysis. *World Journal of Psychiatry*, 7(1), 60-76. <https://doi.org/10.5498/wjpv.7.i1.60>
- Niehaus, K., Fontao, M. I., Garrido Antón, M. J., & Quevedo-Blasco, R. (2025). Characteristics and correlates of cyber-control in Spanish cases of gender-based violence. *Anales de Psicología/Annals of Psychology*, 41(3), 290-302. <https://doi.org/10.6018/analesps.664431>
- Nielsen, M. B., Tangen, T., Idsoe, T., Matthiesen, S. B., & Magerøy, N. (2015). Post-traumatic stress disorder as a consequence of bullying at work and at school. A literature review and meta-analysis. *Aggression and Violent Behavior*, 21, 17-24. <https://doi.org/10.1016/j.avb.2015.01.001>
- Novo, M., Redondo, L., Seijo, D., & Arce, R. (2016). Diseño y validación de una escala para la evaluación del sentido de comunidad en grupos académicos virtuales [Design and validation of a scale to assess the sense of community in academic virtual groups]. *Revista de Investigación en Educación*, 14(2), 126-140. <https://revistas.uvigo.es/index.php/reined/article/view/2080/2056>
- Oberlander, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of content-based techniques to distinguish true and fabricated statements: A meta-analysis. *Law and Human Behavior*, 40(4), 440-457. <https://doi.org/10.1037/lhb0000193>
- Olweus, D. (1973). *Hackkycklingar och övervåld. Forskning om skolmobbing*. Almqvist & Wicksell.
- Olweus, D. (1993). *Bullying at school: What we know and what we can do*. Blackwell.
- Olweus, D. (2012). Cyberbullying: An overrated phenomenon? *European Journal of Developmental Psychology*, 9(5), 520-538. <https://doi.org/10.1080/17405629.2012.682358>
- Olweus, D., & Limber, S. P. (2018). Some problems with cyberbullying research. *Current Opinion in Psychology*, 19, 139-143. <http://dx.doi.org/10.1016/j.copsyc.2017.04.012>
- Palena, N., Cavagnis, L., Vitali, D., Pergolizzi, G. R., & Caso, L. (2024). Improving witnesses' recollection by reinforcing the Cognitive Interview: The 5Ws questions as an information-gathering method. *European Journal of Psychology Applied to Legal Context*, 16(2), 61-75. <https://doi.org/10.5093/ejpalc2024a6>
- Pérez, I. E. (2024). *Análisis de la huella de memoria en casos de violencia de género con las categorías de contenido del Sistema de Evaluación Global* [Doctoral dissertation, University of Santiago de Compostela]. <https://minerva.usc.gal/bitstreams/da72a626-8de8-4e3f-8083-702db9ce5e4b/download>
- Pichel, R., Feijóo, S., Isorna, M., Varela, J., & Rial, A. (2022). Analysis of the relationship between school bullying, cyberbullying, and substance use. *Children and Youth Services Review*, 134(1), Article 106369. <https://doi.org/10.1016/j.childyouth.2022.106369>
- Pichel, R., Foody, M., O'Higgins Norman, J., Feijóo, S., Varela, J., & Rial, A. (2021). Bullying, cyberbullying and the overlap: What does age have to do with it? *Sustainability*, 13(15), Article 8527. <https://doi.org/10.3390/su13158527>
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method bias in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879-903. <http://dx.doi.org/10.1037/0021-9010.88.5.879>
- 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/Annals of Psychology*, 35(1), 156-165. <https://dx.doi.org/10.6018/analesps.35.1.338381>
- Rial-Boubeta, A., Theotonio, Á., Neira-de Paz, A., Braña-Tobío, T., & Varela-Mallou, J. (2024). Relación entre el consumo de videojuegos PEGI18 con violencia explícita, bullying y cyberbullying. *Psychology, Society & Education*, 16(1), 10-19. <https://doi.org/10.21071/pse.v16i1.16718>
- Rogers, R. (2018). An introduction to response styles. In R. Rogers & S. D. Bender (Eds.), *Clinical assessment of malingering and deception* (4th ed.) (pp. 3-17). The Guilford Press.
- Sanmarco, J., Selaya, A., Marcos, V., & Montes, Á. (2023). Contrasting the efficacy of the content analysis categories of the Global Evaluation System to discern between experienced memories and video observed memories. *Revista Iberoamericana de Psicología y Salud*, 14(2), 55-67. <https://doi.org/10.23923/j.rips.2023.02.066>
- Sentencia 213/2002 del Tribunal Supremo, Sala de lo Penal, de 14 de Febrero de 2002. <https://www.poderjudicial.es/search/TS/>
- Sentencia 10554/2014 del Tribunal Supremo, Sala de lo Penal, de 11 de Diciembre de 2014. <https://www.poderjudicial.es/search/TS/openDocument/b6172f0a65e4ba04/20150206>
- Smith, P. K. (2016). Bullying: Definition, types, causes, consequences and intervention. *Social and Personality Psychology Compass*, 10(9), 519-532. <https://doi.org/10.1111/spc3.12266>
- Smith, P. K., del Barrio, C., & Tokunaga, R. (2012). Definitions of bullying and cyberbullying: how useful are the terms? In S. Bauman, J. D. Cross, & J. Walker (Eds.), *Principles of cyberbullying research: Definitions, measures, and methods* (pp. 26-40). Routledge.
- Solberg, M. E., & Olweus, D. (2003). Prevalence estimation of school bullying with the Olweus Bully/Victim Questionnaire. *Aggressive Behavior*, 29(3), 239-268. <https://doi.org/10.1002/ab.10047>

- Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 135-154). Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-94-015-7856-1_8
- Steller, M., & Köhnken, G. (1989). Criteria-based statement analysis. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence*. Springer-Verlag.
- Thomsen, E., Henderson, M., Moore, A., Price, N., & McGarrah, M.W. (2024). *Student reports of bullying: Results from the 2022 school crime supplement to the National Crime Victimization Survey (NCES 2024-109)*. National Center for Education Statistics. <https://nces.ed.gov/pubs2024/2024109rev.pdf>
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327-352. <https://doi.org/10.1037/0033-295X.84.4.327>
- Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Zeugenaussagen [Assessing the credibility of witness statements]. In U. Undeutsch (Ed.), *Handbuch der psychologie, Bd.11: Forensic psychologie* (pp. 26-181). Hogefre.
- Undeutsch, U. (1989). The development of Statement Reality Analysis. In J. C. Yuille Ed. *Credibility assessment* (pp. 101-119). Springer. https://doi.org/10.1007/978-94-015-7856-1_6
- United Nations (UN, 1985). *Declaration of basic principles of justice for victims of crime and abuse of power* [Resolution 40/34]. General Assembly <https://www.unodc.org/pdf/rddb/CCPCJ/1985/A-RES-40-34.pdf>
- United Nations (UN, 2018, July 30). *Protecting children from bullying Report of the Secretary-General (A/73/265)*. <https://docs.un.org/en/A/73/265>
- United Nations Educational, Scientific and Cultural Organization (UNESCO, 2019). *Behind the numbers: Ending school violence and bullying*. United Nations Educational, Scientific and Cultural Organization. <https://unesdoc.unesco.org/ark:/48223/pf0000366483>
- van Der Kolk, B., Ford, J. D., & Spinazzola, J. (2019). Comorbidity of developmental trauma disorder (DTD) and post-traumatic stress disorder: Findings from the DTD field trial. *European Journal of Psychotraumatology*, 10(1), Article 1562841. <https://doi.org/10.1080/20008198.2018.1562841>
- Vilariño, M. (2010). *¿Es posible discriminar declaraciones reales de imaginadas y huella psíquica real de simulada en casos de violencia de género?* [Doctoral dissertation, University of Santiago de Compostela]. <https://minerva.usc.gal/entities/publication/dbc6984a-00c5-4859-93f7-4c3007651106>
- Vilariño, M., Amado, B. G., Seijo, D., Selaya, A., & Arce, R. (2022). Consequences of child maltreatment victimization in internalizing and externalizing mental health problems. *Legal and Criminological Psychology*, 27(2), 182-193. <https://doi.org/10.1111/lcrp.12212>
- Vilariño, M., Novo, M., & Seijo, D. (2011). Estudio de la eficacia de las categorías de realidad del testimonio del Sistema de Evaluación Global (SEG) en casos de violencia de género [Study of the efficacy of the testimony reality categories of the Global Evaluation System (GES) in violence against women cases]. *Revista Iberoamericana de Psicología y Salud*, 2(1), 1-26. <https://www.redalyc.org/articulo.oa?id=245116403001>
- Villarquide, M., Cea, B., Montes, A., & Seijo, D. (2025). Meta-analytic review of the effects of suggestion on children's memory: Implications for testimony. *Anales de Psicología/Annals of Psychology*, 41(3), 312-323. <https://doi.org/10.6018/analesps.665361>
- Vivolo-Kantor, A., Martell, B., Holland, K. M., & Westby, R. (2014). A systematic review and content analysis of bullying and cyber-bullying measurement strategies. *Aggression and Violent Behavior*, 19(4), 423-434. <https://doi.org/10.1016/j.avb.2014.06.008>
- Vrij, A., Leal, S., Deeb, H., & Fisher, R. P. (2025). Omission lies: The effect of omitting little or much information on verbal veracity cues. *European Journal of Psychology Applied to Legal Context*, 17(1), 25-37. <https://doi.org/10.5093/ejpalc2025a3>
- Wachendörfer, M. M., & Oeberst, A. (2024). Differences between true and false memories using the Criteria-Based Content Analysis. *Applied Cognitive Psychology*, 38(5), Article e4246. <https://doi.org/10.1002/acp.4246>
- Woo, J., Chang, S., Hong, J., Lee, D., Hahm, B., Cho, S., Park, J., Jeon, H., Seong, S., Park, J., & Kim, B. (2019). The association of childhood experience of peer bullying with DSM-IV psychiatric disorders and suicidality in adults: Results from a nationwide survey in Korea. *Journal of Korean Medical Science*, 34(46), Article e295. <https://doi.org/10.3346/jkms.2019.34.e295>
- Ye, Z., Wu, D., He, X., Ma, Q., Peng, J., Mao, G., Feng, L., & Tong, Y. (2023). Meta-analysis of the relationship between bullying and depressive symptoms in children and adolescents. *BMC Psychiatry*, 23(1), Article 215. <https://doi.org/10.1186/s12888-023-04681-4>
- Zec, S., Soriani, N., Comoretto, R., & Baldi, I. (2017). High agreement and high prevalence: the paradox of Cohen's kappa. *Open Nursing Journal*, 11, Article 211. <https://doi.org/10.2174/1874434601711010211>