

Reliability Associated with the Psychometric Properties of the Gudjonsson Suggestibility Scale (GSS): A Systematic Review and Meta-Analysis

Fiabilidade associada a las propiedades psicométricas de la Escala de Sugestionabilidad de Gudjonsson (GSS): una revisión sistemática y metaanálisis

Propriedades psicométricas da Escala de Sugestionabilidade de Gudjonsson (GSS): uma revisão sistemática e meta-análise

 Raiza Lorena Sandes Sousa¹
 Marina Pereira Gonçalves¹
 Felipe Negreiros dos Santos¹
 João Carlos Alchieri²
 Jânio Luiz Correia Júnior³
 Ricardo de Freitas-Dias³

¹ Universidade Federal do Vale do São Francisco

² Universidade Federal do Rio Grande do Norte

³ Universidade de Pernambuco

Received: 11/10/2025

Accepted: 05/16/2026

Correspondence

Marina Pereira Gonçalves
marina.goncalves@univasf.edu.br

How to cite: Sandes Sousa, R. L., Pereira Gonçalves, M., Negreiros dos Santos, F., Alchieri, J. C., Correia Júnior, J. L., & de Freitas-Dias, R. (2026). Reliability Associated with the Psychometric Properties of the Gudjonsson Suggestibility Scale (GSS): A Systematic Review and Meta-Analysis. *Ciencias Psicológicas*, 20(1), e-4954.
<https://doi.org/10.22235/cp.v20i1.4954>

Funding: This research did not receive any funding or financial support.

Conflict of Interest: The authors declare no conflict of interest.



Abstract: The Gudjonsson Suggestibility Scales (GSS 1 and 2) assess suggestibility in forensic, clinical, and academic contexts. This study aimed to conduct a systematic review and meta-analysis of the psychometric properties of the GSS. Following PRISMA guidelines, searches were conducted across 12 databases, resulting in 18 eligible studies for inclusion in the systematic review. All stages were independently performed by two trained reviewers to ensure reliability. The results showed that the scales were administered in 10 languages, with evidence of validity and reliability, the latter being more frequently reported across studies. Meta-analytic findings indicated that the GSS generally demonstrated evidence of reliability based on Cronbach's alpha, with values classified as moderate (.75–.79) and acceptable (.70–.74), alongside strong correlations in the random-effects model. Subgroup analyses revealed significant differences between GSS 1 and GSS 2, particularly for the Yield factor. Although the findings indicate acceptable reliability indices, further studies are needed to clarify the internal structure of the GSS and strengthen its contributions to investigative interviewing and the psychology of testimony.

Keywords: suggestibility; witnesses; validity; reliability; systematic review

Resumen: Las Escalas de Sugestionabilidad de Gudjonsson (GSS 1 y 2) evalúan la sugestionabilidad en contextos forenses, clínicos y académicos. Este estudio tuvo como objetivo realizar una revisión sistemática y un metaanálisis de las propiedades psicométricas de GSS. Siguiendo las recomendaciones del PRISMA, se efectuó una búsqueda en 12 bases de datos, identificándose 18 artículos elegibles para su inclusión en la revisión sistemática. Todas las etapas fueron realizadas por dos evaluadores independientes y capacitados, con el fin de garantizar una mayor confiabilidad. Los resultados indicaron que las escalas fueron administradas en diez idiomas, encontrándose evidencias de validez y fiabilidad, siendo esta última más frecuentemente reportada en los estudios. Los resultados del metaanálisis mostraron que, en general, la escala GSS presenta evidencias de fiabilidad basadas en el alfa de Cronbach, con valores clasificados como moderados (.75–.79) y aceptables (.70–.74), además de correlaciones fuertes en el modelo de efectos aleatorios. El análisis de subgrupos reveló diferencias significativas entre los tipos de escala GSS 1 y GSS 2, especialmente en el factor Yield. Se concluye que, aunque estos resultados indican índices aceptables de fiabilidad, aún es necesario desarrollar más estudios para esclarecer la estructura interna de la GSS y potenciar su aplicación en la entrevista investigativa y en la psicología del testimonio.

Palabras clave: sugestionabilidad; testigos; fiabilidad; validez; revisión sistemática

Resumo: As Escalas de Sugestibilidade de Gudjonsson (GSS 1 e 2) avaliam a sugestibilidade em contextos forenses, clínicos e acadêmicos. Este estudo objetivou realizar uma revisão sistemática e metanálise das propriedades psicométricas das GSS. Seguindo as diretrizes PRISMA, foram realizadas buscas em 12 bases de dados, resultando em 18 artigos elegíveis para inclusão na revisão sistemática. Todas as fases foram realizadas independentemente por dois revisores treinados para garantir a confiabilidade. Os resultados mostraram que as escalas foram aplicadas em 10 idiomas, com evidências de validade e confiabilidade, sendo esta última relatada com maior frequência entre os estudos. Os resultados da meta-análise indicaram que as GSS geralmente demonstraram evidências de confiabilidade com base no alfa de Cronbach, com valores classificados como moderados (0,75–0,79) e aceitáveis (0,70–0,74), juntamente com fortes correlações no modelo de efeitos aleatórios. As análises de subgrupos revelaram diferenças significativas entre as GSS 1 e 2, particularmente para o fator Yield. Embora os resultados indiquem índices de confiabilidade aceitáveis, são necessários mais estudos para esclarecer a estrutura interna do GSS e fortalecer suas contribuições para a entrevista investigativa e a psicologia do testemunho.

Palavras-chave: sugestionabilidade; testemunhas; fidedignidade validade; revisão sistemática

Legal psychology is the field of psychology that encompasses the application of psychological knowledge to legal matters (Rovinski & Cruz, 2009). As psychologists became increasingly involved in legal settings, research began to examine the reliability of testimony through the study of psychological processes, seeking to understand the internal mechanisms that facilitate or hinder the accuracy of accounts provided by victims, offenders, and witnesses (Ferreira, 2018).

In this context, the study of suggestion processes—defined as a communicable form of ideation or belief that, once accepted, can induce profound changes in mood, thoughts, perceptions, and behaviors—has become increasingly relevant (Parris, 2016). Consequently, suggestibility is a psychological variable responsible for processing this stimulus and ultimately converting it into a response (suggestion is an external stimulus, while suggestibility is an internal process) (Gudjonsson, 2003).

From this perspective, Gudjonsson (2003, p. 344), an Icelandic psychologist, developed his theory and described interrogative suggestibility (IS) as “the degree to which individuals accept and incorporate information during a formal interrogation, thereby affecting their subsequent responses”. The Gudjonsson and Clark (1986) model further establishes the existence of two distinct types of IS: the first concerns the impact that suggestive or leading questions have on an individual's testimony, while the second refers to the extent to which interviewers can modify undesired responses (whether true or not) through negative feedback.

Based on this IS model, Gudjonsson (1984) created the “Gudjonsson Suggestibility Scale”, available in two versions (GSS-1 and GSS-2), designed to objectively and reliably measure IS. The difference between them lies in the content of the narrative and the questioning: the first version consists of a forensic-themed story, in which GSS-1 describes a fictional robbery. In contrast, GSS-2 depicts a couple preventing a young man from having a bicycle accident; therefore, the narrative content of GSS-2 is somewhat simpler than that of GSS1, making it the version more suitable for children and individuals with learning difficulties. However, the format, number of questions, proportion of leading questions, and structure of the negative feedback are identical. The primary difference between them lies in the narrative content and questioning process: the first version consists of a forensic-content narrative, while the second does not, making it more suitable for studies involving children aged seven and older.

The GSS-1 and GSS-2 provide measures of verbal memory recall, both immediate and delayed, as well as IS, which has been shown to be divided into two components. The author named the two components Yield and Shift. Yield refers to the number of incorrect responses an individual provides during the initial questioning, whereas Shift refers to the number of responses that are changed following negative feedback. Yield and Shift are typically summed and referred to as Total Suggestibility. The GSS-1 and GSS-2 also present clinical indicators through objective measures that the author termed Confabulation, which consists of the sum of two components: Fabrications, which occur when new information is added to the story, and Distortions, which occur when the content of the story is altered in some way. Both reflect the presence of memory errors (Gudjonsson et al., 2020; Merckelbach et al., 1998).

Many studies have investigated variables related to IS and utilized the GSS. Among the main variables studied are: intelligence (Clare et al., 1994; Gudjonsson et al., 1995; Singh & Gudjonsson, 1992; Young et al., 2003), age (Cruz & Pinho, 2014; Gudjonsson et al., 2016; Milne et al., 2002; Rowsell & Colloff,

2024; Saraiva & Albuquerque, 2015), memory (Cunha et al., 2007; Raver et al., 2023; Saraiva & Albuquerque, 2015), and personality (Gorassini et al., 2006; Haraldsson, 1985). Thus, the production of evidence with greater reliability and validity becomes crucial. To achieve this goal, it is necessary to refine the methodologies for interviewing and evaluating testimonies (Gava et al., 2013; Souza & Cruz, 2023; Stein et al., 2009).

Despite the growing relevance of IS in legal psychology, there is a scarcity of standardized, objective, and psychometrically robust instruments specifically designed to assess this construct. Although alternative measures exist, they present important conceptual and methodological limitations. For instance, the Multidimensional Iowa Suggestibility Scale (MISS) assesses general susceptibility to social influence, including persuasion and conformity, but does not capture the specific dynamics of interrogative contexts (Kotov et al., 2004). Similarly, hypnotic suggestibility instruments, such as the Harvard Group Scale of Hypnotic Susceptibility and the Stanford Hypnotic Susceptibility Scales, are based on altered states of consciousness, limiting their ecological validity for forensic settings (Shor & Orne, 1962; Weitzenhoffer & Hilgard, 1959). Another example is the Würzburg Suggestibility Test (WST, Bottenberg & Wehner, 1972; Rowsell & Colloff, 2024), although this instrument lacks validation studies. In addition, indirect approaches, such as measures of memory performance, compliance, or social pressure assess isolated components of suggestibility rather than the integrated process observed during interrogations (Loftus, 2005; Roediger & McDermott, 1995).

In this context, the Gudjonsson Suggestibility Scale (GSS) represents a unique and theoretically grounded instrument, as it combines structured questioning, exposure to misleading information, and controlled negative feedback to simulate key elements of real-world interrogative situations (Gudjonsson, 2003). However, despite its widespread use, concerns remain about the consistency and generalizability of its psychometric properties across diverse populations and contexts. In light of the above, the present study aims to conduct a systematic review with meta-analysis of the psychometric properties of the GSS in its two versions (GSS-1 and GSS-2).

Method

Protocol and registration

This systematic review was conducted in compliance with Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations (Chien et al., 2012). The review protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO), under no CRD42023464581.

Information sources and search strategies

The comprehensive literature search included articles published up to February 2025 and listed in the following electronic databases: Scopus, Academic Search Premier, Embase, PubMed, MEDLINE, CINAHL, APA PsycArticles (APA Journals), SPORTDiscus, LILACS, PsycINFO, PePSIC, SciELO, regardless of study type, population, language, participant age, sex, or publication date, and publication date. Studies were searched using two descriptors: “Gudjonsson Suggestibility Scale” and “Gudjonsson Suggestibility Scales”, combined with the Boolean operator OR. Search terms were adapted to the indexing syntax of each database and restricted to title and abstract fields where available. The complete search strategies are presented in Appendix A, Table A1.

Eligibility criteria

Original studies involving human subjects that reported psychometric properties of the GSS 1 and GSS 2 scales (evidence based on content, evidence based on the item response process, evidence based on internal structure, evidence based on relationships with external variables, and reliability or consistency of the measure) were included in this systematic review. Review studies, gray literature, editorials, conference publications, theses/dissertations, unpublished studies in indexed platforms, duplicate records, and studies that mentioned or used the GSS, but without analyzing psychometric characteristics of the GSS (1 or 2), were excluded.

Study selection

The study selection process, along with the evaluation form, was designed based on predefined inclusion and exclusion criteria. This form was calibrated prior to the commencement of the screening

process for study selection. Using EndNote X9 software (Clarivate Analytics), duplicate studies were identified and removed by a trained researcher (JLCJ).

Subsequently, the retrieved studies were exported to the Rayyan QCRI systematic review tool (Qatar Computing Research Institute) (Hupe, 2019; Ouzzani et al., 2016) for record management. RLSS and FNS independently screened potential studies for inclusion, first through title and abstract review, followed by a full-text evaluation. Additionally, studies shortlisted for in-depth reading were cross-referenced by examining their reference lists. This step aimed to identify relevant records that may not have been captured during the electronic database search (Higgins et al., 2019).

Data extraction

Data extraction was performed according to the Cochrane Handbook for Systematic Reviews of Interventions (Higgins et al., 2019). Data extracted from studies satisfying eligibility criteria were entered into an electronic Excel spreadsheet. The following data were extracted from the studies: a) study registration data, b) objective and type of study, c) study methodology, d) characteristics of the instrument, e) results of the psychometric properties, and f) statements (funding and conflicts of interest). Finally, we summarized the observed results of all the psychometric properties mentioned above, subdivided into summary and/or of the results, general result, and certainty of the evidence.

Assessment of the risk of bias

The Critical Appraisal Tool (CAT) for Validity and Reliability Studies of Objective Clinical Tools was utilized to evaluate the methodological quality (Brink & Louw, 2012). The thirteen items in this tool were used for validity and reliability studies (items 1, 2, 10, 12, and 13); validity studies (items 3, 7, 9, and 11); and reliability studies (items 4, 5, 6, and 8).

Prior to study selection, data extraction, and risk-of-bias assessment, raters underwent calibration training to ensure interrater consistency. Disagreements between the two review authors (RLSS and FNS) regarding study selection, data extraction, and risk of bias assessments were resolved through discussion, with the involvement of a third author (MPG) when necessary.

Data analysis

Initially, the psychometric properties of the selected studies were analyzed, followed by a meta-analysis of the quantitative results from the studies included in the review. Specifically, the meta-analysis focused on reliability evidence as an aspect of validity based on internal structure, in accordance with contemporary psychometric frameworks (American Educational Research Association [AERA] et al., 2014).

To aggregate the Cronbach's alpha (α) values for each GSS factor, we applied a variance-stabilizing transformation (Fisher's Z) to the coefficients prior to meta-analysis, as recommended for bounded reliability indices that can be expressed as functions of inter-item correlations (Bonett, 2002; Field & Gillett, 2010). Although Cronbach's alpha is not a simple correlation, this approach is commonly adopted in meta-analytic studies of reliability to approximate normality and stabilize sampling variance across studies. The meta-analyses were calculated using random effects (RE) models, computed through the restricted maximum likelihood method, which assumes that the selected studies were sampled from a larger set of studies and explicitly models between-study heterogeneity (Higgins et al., 2003; Prinsen et al., 2018). We also explicitly recognize the limitations of alpha, including its sensitivity to scale length and assumptions of tau-equivalence, and interpret the pooled estimates with appropriate caution (Dunn et al., 2014; McNeish, 2018).

Note that the random-effects model was chosen over a fixed effect model, because study-level and sample-level characteristics (e.g., educational level, sex, and age) may act as sources of between-study heterogeneity, influencing the magnitude and variability of the reported psychometric estimates across studies. In addition, the random-effects model accounts for both within-study error and between-study variance, allowing for broader generalization of the findings beyond the specific set of included studies, which is recommended when heterogeneity is expected (Borenstein et al., 2009; Higgins et al., 2003).

Statistical heterogeneity was investigated using the Cochran Q test (adopting a significance level of $p < .10$, as recommended in meta-analytic studies due to the low statistical power of this test, particularly when the number of studies is limited). Statistical inconsistency was investigated using the

Higgins I² statistic (Higgins et al., 2003), as follows: $\leq 40\%$, low heterogeneity; 30% to 60% , moderate heterogeneity; $> 50\%$ to 90% , substantial heterogeneity; and $> 75\%$ to 100% , considerable heterogeneity (Higgins et al., 2019). Whenever $I^2 > 50\%$ and tau squared (τ^2) > 1 , in conjunction with statistical significance in the Q test ($p < .10$), heterogeneity was considered substantial and potential sources were further investigated.

For analyzing internal consistency and qualitative interpretation, we employed Cronbach's alpha values, using interpretative benchmarks adapted from the literature as general guidelines rather than strict cut-off criteria, adopting the following categories: excellent $\geq .85$; good $.80-.84$; moderate $.75-.79$; and fair $.70-.74$. It is important to note that the interpretation of Cronbach's alpha is context-dependent and may vary according to the number of items, construct complexity, and measurement purpose. Cronbach's alpha values were interpreted according to factor structure and sample size (Ponterotto & Ruckdeschel, 2007; Santos et al., 2020).

For results regarding evidence based on internal structure, the following criteria were considered acceptable following the literature: Comparative Fit Index (CFI) or Tucker–Lewis Index (TLI) $> .95$; Root Mean Square Error of Approximation (RMSEA) $< .06$; Standardized Root Mean Square Residual (SRMR) $< .082$; and Average Variance Extracted (AVE) $> .50$ (Hair et al., 2021; Prinsen et al., 2018; Terwee et al., 2007), these indicators being qualitatively evaluated in the studies included in the systematic review. Regarding evidence based on relations with external variables (e.g., convergent, discriminant, concurrent, or predictive validity), this evidence was not quantitatively synthesized but assessed qualitatively due to heterogeneity and limited availability of comparable data across studies. For this purpose, coefficients $> .70$ were used as interpretative benchmarks, in line with established recommendations for the qualitative evaluation of measurement properties (Prinsen et al., 2018). For reliability, assessed using Intraclass Correlation Coefficients (ICC) and composite reliability, values $> .70$ were also deemed acceptable (Prinsen et al., 2018).

Subgroup analysis

Subgroup analyses were conducted to explore potential sources of heterogeneity across studies. Specifically, analyses were performed according to the type of GSS instrument (GSS 1 vs. GSS 2), as these differ in the content of the narrative and the questioning, with GSS1 consisting of a narrative with forensic content and GSS2 depicting a couple preventing a young man from having a bicycle accident, being the most suitable for studies with children from 7 years of age, for example. Thus, as such differences may introduce measurement-related variability in psychometric properties across studies (Borenstein et al., 2009; Gudjonsson, 1997, 2003), the subgroup analysis by scale type is justified.

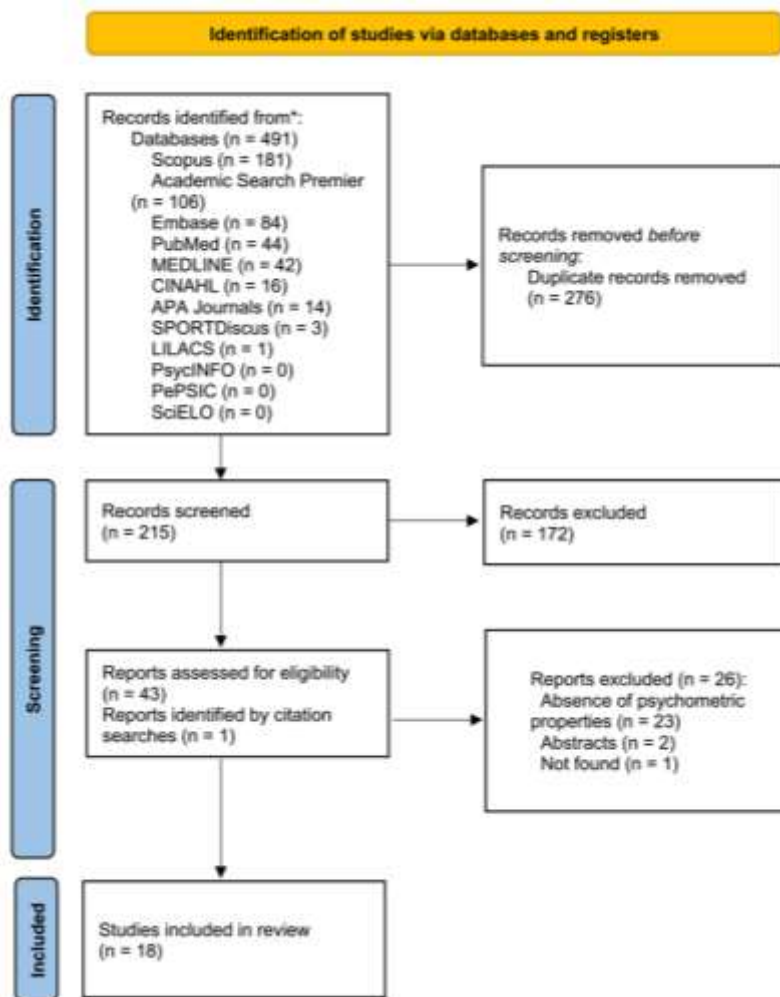
All analyses were performed using the “metacor” (Meta-analysis of single correlations) function of the meta package version 5.2–0 (Balduzzi et al., 2019), implemented in the statistical program R version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria), and followed the recommendations of the Cochrane Handbook for Systematic Reviews and the guideline COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) (Higgins et al., 2019; Mokkink et al., 2018).

Results

A total of 491 articles were found in the selected databases. Of these, 276 duplicates were excluded. Title/abstract screening and full-text review included 215 and 43 articles, respectively, with 74.2% agreement between raters. Eighteen of these articles satisfied eligibility criteria and were included in the quantitative narrative analysis of this meta-analysis (Figure 1). After the cross-referencing search, one article was identified that met the eligibility criteria; however, we were unable to access these articles. Despite requesting it from the authors, we did not receive a response before the conclusion of this review.

Figure 1

Flowchart of the selection of studies



Characteristics of the studies

For the narrative synthesis of this systematic review, studies published between 1984 and 2024 were included. The studies were published in English and Spanish. The GSS 1 and 2 scales were administered in 10 languages: English, Portuguese, Polish, Spanish, Turkish, Italian, German, Dutch, Swedish, and Japanese. The total sample consisted of 4631 participants, with sample sizes ranging from 31 to 1,261 participants per study. Of these, 2144 were women (46.30 %), 2303 were men (49.73 %), 181 were from studies that did not specify gender (3.91 %), and 3 participants preferred not to disclose their gender (0.06 %). Regarding age, participants ranged from 6 to 78 years, with 15 studies involving adults and 2 studies involving children and adolescents.

Among the studies analyzed, 14 were cross-sectional observational studies, 3 were longitudinal (with intervals between GSS applications ranging from 1 day to 4 months), and 2 followed an experimental design. Regarding the scales, 4 studies collected data from both GSS 1 and GSS 2, while 8 studies focused on GSS 1 and 6 on GSS 2. The psychometric evidence identified in the selected articles includes evidence of reliability and validity, internal structure, relationships with external variables, and content. However, the other psychometric properties outlined by AERA et al. (2014) were not identified. A detailed description of the characteristics of the included studies is provided in Table 1. Table 2 presents the synthesis and quantitative results concerning the psychometric validity properties of the GSS. Table A2 (Appendix A) provides the list of excluded studies, all of which were excluded due to the absence of GSS psychometric properties.

Table 1

Synthesis and characteristics of the results of the studies of the instrument GSS

Authors and year	Type of study	Sample size	Characteristics of participants	Validation Language	Application	Types of psychometric properties	Competing Interest (CI) / Funding (F)
Angelucci, 2005	Cross-sectional observational study	N = 320	Undergraduate students from the UCAB. 143 female and 177 male, age range 16–28 years, mean age 19.48.	Spanish (Venezuela) GSS1	In-person	EBI; RE	CI – ND; F – ND
Bianco & Curci, 2015	Study 1 Cross-sectional observational study	N = 1,261	Group 1: 624 participants (52.6 % women), with a mean age of 37.54 ($\pm$ 13.91; range = 16–78). Group 2: 637 participants (58.4 % women) with a mean age of 36.67 ($\pm$ 14.29; range = 16–84).	Italian GSS1 and GSS2	In-person	RE; EBI	CI – ND; F – ND
	Study 2 Cross-sectional observational study	N = 102	Group 1: 50 participants (52 % women), with a mean age of 37.66 ($\pm$ 13.74). Group 2: 52 participants (51.9 % women) with a mean age of 36.27 ($\pm$ 14.14).	Italian GSS1 and GSS2	In-person	RE; EBE	CI – ND; F – ND
Clare et al., 1994	Cross-sectional observational study	N = 101	101 participants aged between 17 and 69 years (M = 33.2; $\pm$ 10.5).	English (United Kingdom) GSS2	In-person	RE	CI – ND; F – ND
Gignac & Powell, 2009	Cross-sectional observational study	N = 220	220 children aged 5.3–13.2 years (M = 10.1, $\pm$ 2.16; 43.2 % female).	English (Australia) GSS2	NR	RE; EBI	CI – ND; F – ND
Gudjonsson, 1984	Cross-sectional observational study	N = 195	195 participants. These comprised four independent groups: (i) 58 'normal' females, mean age of 31 years ($\pm$ 12.7); (ii) 56 'normal' males, mean age 24 years ($\pm$ 5.8); (iii) 40 forensic patients mean age 30 years ($\pm$ 9.1); (iv) 41 children (mainly delinquent boys), mean age 14 years ($\pm$ 1.4).	English (United Kingdom) GSS1	In-person	RE	CI – ND; F – ND

Authors and year	Type of study	Sample size	Characteristics of participants	Validation Language	Application	Types of psychometric properties	Competing Interest (CI) / Funding (F)
Gudjonsson, 1987	Longitudinal observational study	<i>N</i> = 90	Three groups of participants were tested: (1) Twenty-eight 'normal' males (<i>n</i> = 14) and females (<i>n</i> = 14) with a mean age of 27 years ($\pm$ 7.7). The majority of the subjects were hospital staff. (2) Thirty-two 'forensic' patients and 'clients' referred to the author by defense or prosecution solicitors for court reports. There were 27 males and five females in the group, having a mean age of 31 years ($\pm$ 11.9). (3) Thirty 'forensic' cases similar to those in group 2, with a mean age of 31 years ($\pm$ 10.3). There were 28 males and two females in this group.	English GSS1 and GSS2	In-person	RE	CI – ND; F – ND
Gudjonsson, 1992	Cross-sectional observational study	<i>N</i> = 129	129 participants; 100 were forensic male patients, with a mean age of 28 ($\pm$ 10), and 29 were normal adult males, with a mean age of 26 ($\pm$ 5).	English (United Kingdom) GSS1 and GSS2	In-person	RE	CI – ND; F – ND
Gudjonsson & Singh, 1984	Cross-sectional observational study	<i>N</i> = 31	31 boys with a mean age of 14.3 yr ($\pm$ 1.3, range 11–16).	English (United Kingdom) GSS1	In-person	RE	CI – ND; F – ND
Hein et al., 2021	Cross-sectional observational study	<i>N</i> = 560	<i>N</i> = 560 participants (<i>n</i> = 287 female; <i>n</i> = 283 from the University of Kiel; <i>n</i> = 277 from the University of Bonn). Age ranged from 18 to 60 years (<i>M</i> = 24.20 $\pm$ 4.60, median = 23).	German GSS1	Online	EBI; RE	CI – ND; F – Funded by German Research Foundation (BE 2443/11–1, LE 2240/6-1)

Authors and year	Type of study	Sample size	Characteristics of participants	Validation Language	Application	Types of psychometric properties	Competing Interest (CI) / Funding (F)
Leue & Bertram, 2024	Study 1 Cross-sectional observational study	N = 206	N = 206 participants, 121 participants were male, 85 were females. Participants were aged 18 to 41 years ($M = 23.87, \pm 3.64$).	German GSS1	Online	EBI; RE	CI – The authors declare that there were no conflicts of interest; F – Funded by German Research Foundation
	Study 2 Cross-sectional observational study	N = 101	N = 101 participants (78 female, 23 male, 1 no information for gender). Participants were 18 to 61 years old ($M = 24.85, \pm 7.34$).	German GSS1	Online	EBE; RE	CI – The authors declare that there were no conflicts of interest; F – Funded by University of Kiel
Merckelbach et al., 1998	Study 1 Longitudinal observational study	N = 40	40 undergraduate students (37 women and 3 men; mean age = 20.6 years, ± 2.2 , range 18–26).	Dutch GSS1 and GSS2	In-person	RE	CI – ND; F – ND
	Study 2 Experimental	N = 53	53 undergraduate students (41 women and 12 men; mean age = 20.0 years; range 18–27).	Dutch GSS1 and GSS2	In-person	EBE	CI – ND; F – ND
Özcan & Ziyalar, 2023	Cross-sectional observational study	N = 175	The ages of the participants ranged from 19 to 36. 104 were female, 68 were male, and 3 people did not specify their gender. Most of the participants were undergraduate students. The average age of the participants was 23.3.	Turkish GSS2	In-person	RE	CI – The authors declare that there were no conflicts of interest; F – The authors stated that this study did not receive financial support
Polczyk, 2005	Cross-sectional observational study	N = 148	Sample 1 consisted of 148 participants, 89 female and 59 males with a mean age of 22.8, ± 4.0 (range 18–56).	Polish GSS1	In-person	RE	CI – ND; F – ND
	Cross-sectional observational study	N = 207	Sample 2 comprised 207 participants (120 female and 87 males; mean age 22.4, ± 3.5)	Polish GSS2	In-person	RE	CI – ND; F – ND
Polczyk, 2016	Cross-sectional observational study	N = 118	118 participants (69 women and 49 men), with mean age 22.20 years (± 2.00 , range 18–29).	Polish GSS2	In-person	RE	CI – ND; F – ND

Authors and year	Type of study	Sample size	Characteristics of participants	Validation Language	Application	Types of psychometric properties	Competing Interest (CI) / Funding (F)
Richardson & Smith, 1993	Cross-sectional observational study	<i>N</i> = 57	57 young people (48 male, 9 female) in a specialized facility, all exhibiting disturbed and disordered behaviour, including criminal offences. The age range was 10.9–17.6 years (<i>M</i> = 15.5).	English (United Kingdom) GSS1	In-person	RE	CI – ND; F – ND
Singh & Gudjonsson, 1987	Cross-sectional observational study	<i>N</i> = 285	285 GSS forms obtained from clinical practice and previous research. 209 (73 %) of the participants were male and 76 (27 %) were female. The age range was 11–65 years.	English (United Kingdom) GSS1	In-person	RE	CI – ND; F – ND
Wachi et al., 2019a	Longitudinal observational study	<i>N</i> = 442	<i>N</i> = 442. 218 (49.3 %) males and 224 (50.7 %) females). Mean and median ages were 40.12 ($\pm$ 10.66) and 40.00 years, respectively (range: 20–59 years).	Japanese GSS2	Online and In-person	CBE; RE	CI – ND; F – ND
Wachi et al., 2019b	Experimental Study	<i>N</i> = 218	218 people (male = 108, 49.5 %; female = 110, 50.5 %) who engaged in both immediate recall on the first day and delayed recall two days later. The mean age was 39.87 ($\pm$ 10.71) with a range from 20 to 59.	Japanese GSS2	Online	RE	CI – ND; F – ND

Note. CBE: Content-Based Evidence; EBI: Evidence Based on Internal Structure; EBE: Evidence Based on Relations to External Variables; RE: Reliability Evidence; NR: Not Reported; ND: Not Declared.

Table 2

Synthesis and quantitative results of the psychometric properties of validity of the studies of the instrument GSS

Authors and year	Evidence based on internal structure	Evidence based on relationships with external variables	Reliability
Angelucci, 2005	—	—	Cronbach's Alpha: Yield = .62; Shift = .47
Bianco & Curci, 2015	RMSEA < .10; SRMR < .08	—	Kuder-Richardson's coefficient GSS1: Yield 1 = .66; Yield 2 = .71; Shift = .58; Total suggestibility = .68. Kuder-Richardson's coefficient GSS2: Yield 1 = .72; Yield 2 = .77; Shift = .53; Total suggestibility = .73.
	3 factors: RMSEA = .07; AIC = 7763.75; SRMR = .08	—	—
Clare et al., 1994	—	—	Inter-rater reliability: Yield 1 = .99; Yield 2 = .99; Shift = .99; Total Suggestibility = .99 ($p < .001$).
Gignac & Powell, 2009	—	—	Cronbach's Alpha: Yield = .84; Shift = .67. McDonald's Omega: Yield 1 = .83; Shift = .60.
Gudjonsson, 1984	—	—	Cronbach's Alpha: Yield = .77; Shift = .67.
Gudjonsson, 1987	—	—	Test-retest scores: Yield 1 = .84; Yield 2 = .86; Shift = .79; Total suggestibility = .90.
Gudjonsson, 1992	—	—	Cronbach's Alpha (GSS1): Yield 1 = .87; Yield 2 = .90; Shift = .79.
Gudjonsson & Singh, 1984	—	—	Cronbach's Alpha: Yield = .77; Shift = .69.
Hein et al., 2021	2 factors: Yield 1 [$\chi^2 = 174.43$, $df = 151$, $p = .09$]; RMSEA = .02; CFI = .89]; Yield 2 [$\chi^2 = 184.02$, $df = 149$, $p < .05$]; RMSEA = .02; CFI = .94]. 1 factor: Shift [$\chi^2 = 181.74$, $df = 170$, $p = .26$; RMSEA = .01; CFI = .98]]	—	Cronbach's Alpha: Yield = .43; Shift = .66. Spearman-Brown corrected split-half coefficients: Yield 1 = .45; Yield 2 = .71; Shift = .68. Hancock's H: Yield 1 = .59; Yield 2 = .71; Shift = .71.
Leue & Bertram, 2024	2 factors: RMSEA = 0.00; CFI = 1.00; SRMR = 0.05	—	Cronbach's Alpha: immediate recall = .89. Split-half reliability (1st vs. 2nd half, Spearman-Brown corrected): immediate recall = .72. Split-half reliability (odd-even method, Spearman-Brown corrected): immediate recall = .88.
	—	Yield 1, Yield 2, and Shift did not significantly correlate with I-S-T 2000 R ($rs(101) = .09$ to $.17$)	—

Merckelbach et al., 1998	—	—	Cronbach's Alpha: Yield = .79; Shift = .75; Total = .82. Test-retest: $r = .55$ ($p < .01$).
	—	Yield and DES correlation: $r = .40$ ($p < .01$)	—
Özcan & Ziyalar, 2023	$\chi^2/df = 0.05$; RMSEA = .00; GFI = .99; SRMR = .016	—	Cronbach's Alpha: Yield 1 = .713; Yield 2 = .812; Shift = .600; Total suggestibility = .74.
Polczyk, 2005	—	—	Cronbach's Alpha: Yield 1 = .74; Yield 2 = .82; Shift = .63; Total suggestibility = .77.
	—	—	Cronbach's Alpha: Yield 1 = .68; Yield 2 = .71; Shift = .63; Total suggestibility = .71.
Polczyk, 2016	—	—	Cronbach's Alpha: Yield = .76; Shift = .66; Total Suggestibility = .78.
Richardson & Smith, 1993	—	—	Inter-rater reliability: Yield 1 $r = .983$ ($p < .001$); Yield 2 $r = .994$ ($p < .001$); Shift $r = .949$ ($p < .001$); Total score $r = .992$ ($p < .001$).
Singh & Gudjonsson, 1987	—	—	Cronbach's Alpha: Yield = .701; Yield 2 = .713; Shift = .661; Total suggestibility = .665.
Wachi et al., 2019a	—	—	Kappa: Yield 1 = .968; Yield 2 = .989; Shift = .969. Inter-rater reliability: $r = 0.997$.
Wachi et al., 2019b	—	—	Inter-rater reliabilities: Yield 1 = 99.1 % (kappa = .97); Yield 2 = 99.4 % (kappa = .99); Shift = 98.9 % (kappa = .97).

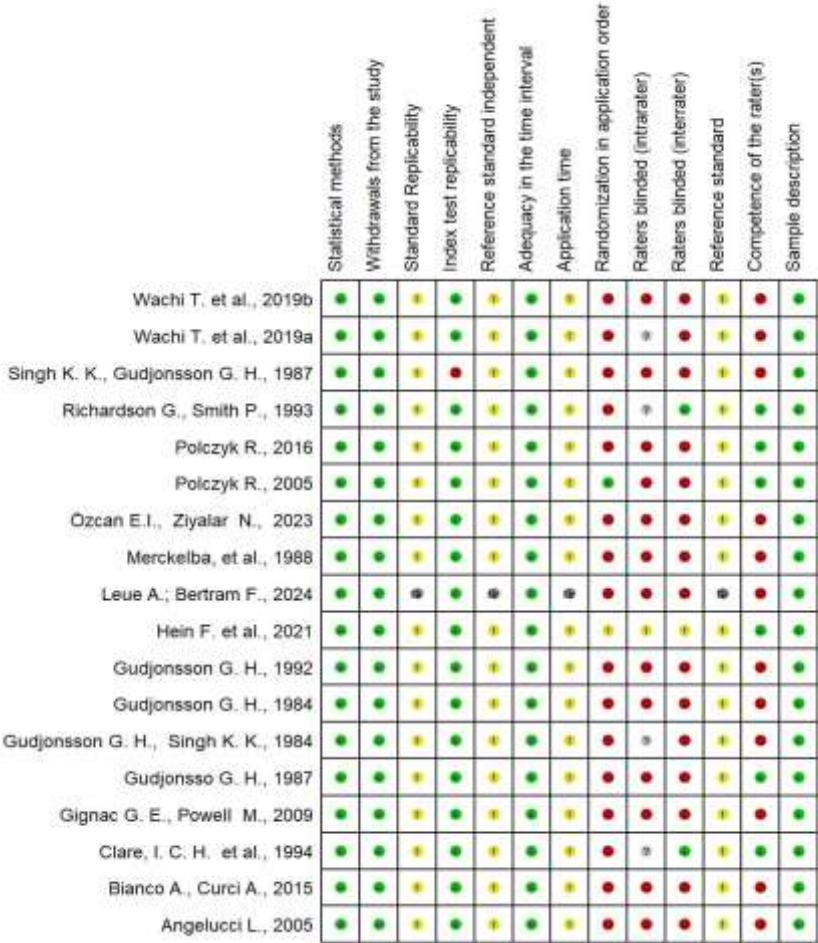
Note. Cronbach's alpha values, assessed according to the following categories: excellent ($\geq .85$); good (.80–.84); moderate (.75–.79) and fair (.70–.74). For the results of evidence based on internal structure, the following criteria were considered acceptable: Comparative Fit Index (CFI) or Tucker–Lewis Index (TLI) $> .95$; Root-Mean-Square Error of Approximation (RMSEA) $< .06$; Standardized Root Mean Square Residual (SRMR) $< .082$; Average Variance Extracted (AVE) $> .50$. For the results of evidence based on relationships with external variables (criterion validity, convergent or discriminant type; and concurrent or predictive type), results $> .70$ were considered acceptable. For reliability results, assessed by Intraclass Correlation Coefficients (ICC) and composite reliability, results $> .70$ were considered acceptable (Prinsen et al., 2018).

Risk of bias analysis

For this stage of the analysis, the Critical Appraisal Tool for Validity and Reliability Studies of Objective Clinical Tools (Brink & Louw, 2012) was employed. It is important to note that, during the evaluation of the studies, the designation “NR” (not reported) was applied to certain procedures specifically related to the use of a gold standard test for measuring the assessed construct—namely, criteria 03, 07, 09, and 11—since, to date, no standardized measurement of this construct has been established. Conversely, the designation “NA” (not applicable) is already provided within the tool itself, consistent with the validity standards developed and tested by Brink and Louw (2012). When evaluating the validity and reliability criteria in the reviewed studies, it was observed that all studies met criteria 01, 08, 10, 12, and 13, with the exception of the study by Singh and Gudjonsson (1987). Criterion 02, in contrast, received a positive rating in only six studies (Clare et al., 1994; Gudjonsson, 1987; Hein et al., 2021; Polczyk, 2005, 2016; Richardson & Smith, 1993).

Regarding the criteria assessing reliability, only two studies reported inter-rater blinding (criterion 04), namely Richardson and Smith, (1993) and Clare et al., (1994). Intra-rater blinding (criterion 05) was not achieved in any of the evaluated studies, while criterion 06 was met only in the study by Polczyk (2005). Taken together, criteria 04, 05, and 06 highlight methodological weaknesses that may introduce a risk of bias into the systematic review (Figure 2). The high frequency of NA and NR may suggest that many studies did not meet critical validity and reliability criteria, indicating either a lack of adherence to rigorous methodological standards or that certain methodological aspects were not considered important or relevant for some studies, which may also limit the applicability of the results.

Figure 2
Methodological Quality Assessment of the Included Studies



Note. Green circles indicate low risk of bias (criterion fully satisfied); Yellow circles indicate unclear or moderate risk of bias (insufficient information or partial compliance); Red circles indicate high risk of bias (criterion not satisfied); and Grey circles (with “?”) indicate missing or not reported data.

Synthesis of reliability evidence through meta-analysis

Reliability evidence for the GSS scale was assessed across three factors: Yields, Shift, and Total. For the Yields factor, 10 studies totaling 1,823 observations were included, with a random-effects model yielding a Cronbach’s alpha (α) of .75 (95 % CI [.67, .81], $p < .001$). Heterogeneity was high ($I^2 = 94\%$), indicating substantial variability among studies (Figure 3). For the Shift factor, 11 studies with 2,108 observations were included. The random-effects model yielded $\alpha = .67$ (95 % CI [.64, .70], $p < .001$). Heterogeneity was lower compared to the Yields factor, with $I^2 = 36.5\%$, indicating moderate variability among studies (Figure 4). For the Total factor, 7 studies with 1,129 observations were analyzed. The random-effects model showed $\alpha = .82$ (95 % CI [.63, .91], $p < .001$). Heterogeneity was very high, with $I^2 = 97.5\%$, indicating substantial variability across studies (Figure 5).

Figure 3
Comparison of Internal Consistency Evidence for the Yields Factor Across GSS Scale Studies Using the Bifactor Model

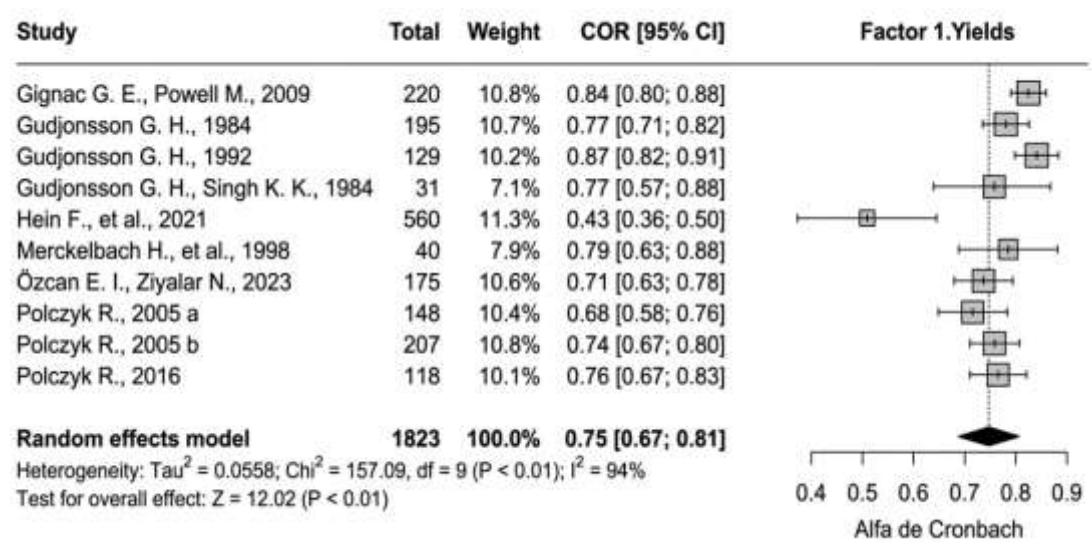


Figure 4
Comparison of Internal Consistency Evidence for the Shift Factor Across GSS Scale Studies Using the Bifactor Model

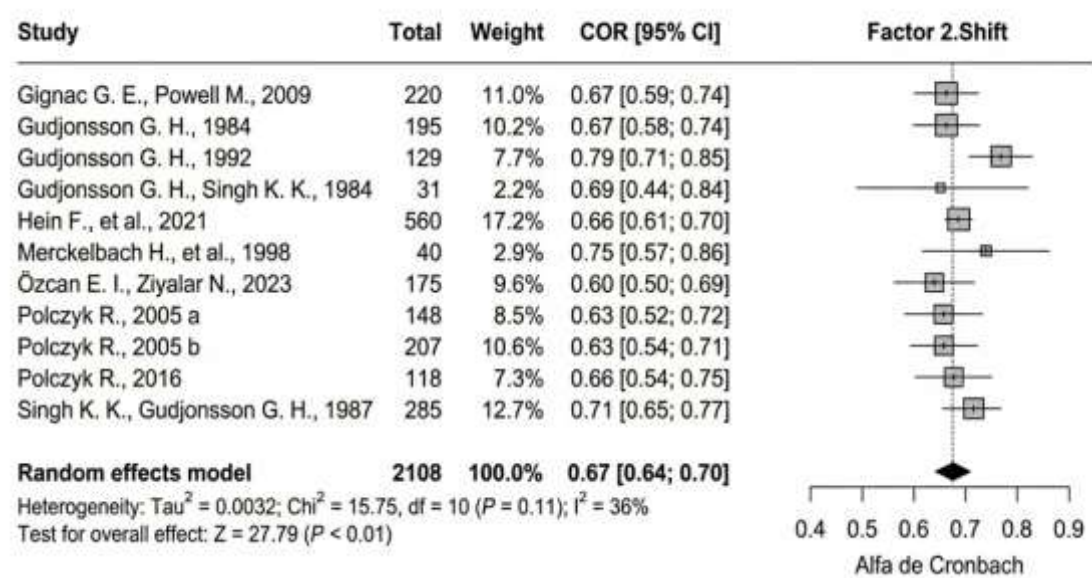
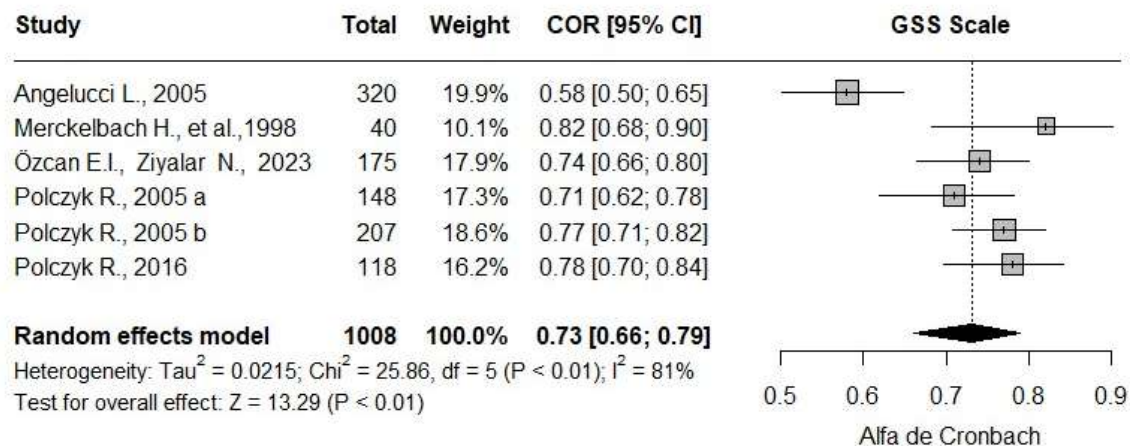


Figure 5

Comparison of Internal Consistency Evidence for the Total Score Across GSS Scale Studies Using the Bifactor Model



Subgroup Analysis

In the subgroup analysis for the Yields factor, comparing GSS 1 and GSS 2 scale types, GSS 2 showed $\alpha = .79$ (95 % CI [.73, .84], $I^2 = 78\%$), while GSS 1 showed $\alpha = .67$ (95 % CI [.49, .80], $I^2 = 95\%$). The difference between groups was statistically significant ($p < .001$). For the Shift factor, GSS 2 had $\alpha = .68$ (95 % CI [.61, .74], $I^2 = 61\%$), whereas GSS 1 had $\alpha = .67$ (95 % CI [.64, .70], $I^2 = 0.0\%$), with no significant difference between groups ($p = .9861$). For the Total factor, GSS 2 showed $\alpha = .77$ (95 % CI [.73, .80], $I^2 = 0.0\%$), while GSS 1 showed $\alpha = .86$ (95 % CI [.33, .98], $I^2 = 99\%$). No significant difference was observed between groups ($p = .5199$) (Figures 3, 4, and 5).

In summary, the results indicate that, overall, the GSS scale demonstrates reliability evidence based on Cronbach's alpha, with values classified as moderate (.75–.79) and fair (.70–.74) (Ponterotto & Ruckdeschel, 2007; Santos et al., 2020), showing strong correlations in the random-effects model. However, for the Yields factor, one study (Hein et al., 2021) reported a Cronbach's alpha below the acceptable threshold ($\alpha = .43$) with high heterogeneity for this factor ($I^2 = 94\%$). For the Shift factor, the meta-analyzed studies showed an alpha slightly below the recommended level [$\alpha = .67$; 95 % CI [.65, .69], $p < .001$], with low heterogeneity among studies ($I^2 = 36.5\%$). Heterogeneity therefore varied across factors, being highest for the Yields and Total factors. Subgroup analysis revealed significant differences between GSS 1 and GSS 2 scale types, particularly for the Yields factor.

Discussion

This systematic review analyzed studies that presented the psychometric properties of the GSS 1 and GSS 2 scales. Regarding Content Validity Evidence (CVE), this type of validity was identified in the study by Özcan and Ziyalar (2023). The authors reported that the translation and back-translation of the GSS form were conducted by four psychologists fluent in English. The back-translations were compared with the original form, and the most similar translation was selected for the study. A similar type of validity evidence was observed in the study by Angelucci (2005), which analyzed the Spanish version of the GSS. The study reported that the adaptation process began with the original version of the GSS 1 by Gudjonsson (1984), translated by the author herself. The translation was reviewed and corrected by three judges. Additionally, modifications were made to adapt the questions and narrative to the Venezuelan context.

Polczyk (2005) also described an adaptation. According to the author, the scales were carefully translated from English to Polish to ensure linguistic equivalence. Some details in the narrative, such as geographical names, were modified to align with Polish cultural and contextual realities. However, no back-translation was performed, as the author argued that minor changes to the story's content or questions would not affect the overall concept of the scales.

Thus, the adaptation processes for Turkish, Spanish, and Polish described by Özcan and Ziyalar (2023), Angelucci (2005), and Polczyk (2005) only partially followed the steps for content validation

considered appropriate in the literature. This partial adherence may somewhat compromise the assessment of IS in these countries. Other studies on the adaptation and validation of the GSS scales in different languages did not meet the inclusion criteria or were not analyzed due to lack of full access to the publications.

Regarding Validity Evidence Based on Internal Structure (EBI), several studies conducted Confirmatory Factor Analyses (CFA), totaling seven publications (Angelucci, 2005; Bianco & Curci, 2015; Gignac & Powell, 2009; Gudjonsson, 1984, 1992; Hein et al., 2021; Özcan and Ziyalar, 2023).

The first indications of the factorial structure of the GSS 1 and GSS 2 scales were presented by their creator, Gudjonsson (1984, 1992), in the 1980s. The author conducted Principal Component Analyses (PCA) with Varimax rotation, using a sample of 195 participants. Two suggestibility factors emerged: (1) Yield 1 (compliance) and (2) Shift (change). The identified factorial structure was theoretically compatible with Gudjonsson's proposed model for explaining IS.

In a subsequent study in the 1990s, Gudjonsson (1992) applied the GSS 2 to a sample of 129 male participants. A factor analysis with Varimax rotation was performed, using the same statistical procedure previously applied to the GSS 1 (Gudjonsson, 1984). The results confirmed that the Yield 1 and Shift scores loaded onto separate factors, supporting the theoretical framework proposed by Gudjonsson and Clark (1986).

Subsequent studies have tested the factorial structure of the GSS 1 and GSS 2 scales in different contexts and with diverse samples. For instance, Bianco and Curci (2015) investigated the Italian adaptation and identified a three-factor structure (Yield 1, Yield 2, and Shift) rather than the original bifactorial structure (Yield 1 and Shift). The fit indices used in their study included Chi-square/*df*, RMSEA, SRMR, and AIC, which were considered appropriate based on psychometric guidelines (Hair et al., 2021; Prinsen et al., 2018; Terwee et al., 2007). Their analysis revealed that the three-factor solution provided a better model fit to the data ($\chi^2/df = 4.40$, RMSEA = .07, AIC = 7763.75, SRMR = .08). However, the differences in fit indices among the three models were minimal or non-significant ($\chi^2/df = 6.16$, n.s.; $\chi^2/df = 6.03$, n.s.) (Bianco & Curci, 2015).

Gignac and Powell (2009) analyzed the GSS 2 and questioned the bifactorial model. According to them, the Shift subscale should be considered a difference score index, as it represents the difference between responses to Yield 1 and Yield 2 items. To test this hypothesis, the authors conducted a CFA with a sample of 220 Australian children. Their results indicated that Yield 1 scores reliably measured a construct; however, the Shift subscale exhibited an inadequate level of internal consistency reliability (McDonald's omega = .60). Due to this limitation, the authors recommended cautious use of the GSS scales and advised against using the Shift subscale until further validity studies are conducted.

Hein et al. (2021) conducted a CFA with a sample of 560 German students who responded to the GSS 1. The bifactorial model showed an acceptable fit for Yield 1 [$\chi^2 = 174.43$, *df* = 151, *p* = .09]; RMSEA = .02; CFI = .89] and for Yield 2 [$\chi^2 = 184.02$, *df* = 149, *p* < .05]; RMSEA = .02; CFI = .94]. Given that the acceptance of misinformation suggested by the key questions is reflected in the Yield 2 score, the results supported the construct's theoretical basis. Additionally, the single-factor model for Shift demonstrated a very good model fit [$\chi^2 = 181.74$, *df* = 170, *p* = 0.26; RMSEA = .01; CFI = .98], further confirming the theoretical foundation of the German online version of the GSS 1.

Another study that provided evidence of internal structure validity through CFA was Özcan and Ziyalar (2023), which assessed the GSS 2 in the Turkish context, with a sample of 175 undergraduate students. The analysis yielded the following fit indices: ($\chi^2/df = 0.05$, RMSEA = .001, GFI = .99, SRMR = .016). However, the Shift factor did not exhibit unidimensional properties, as some items did not align with the factor structure.

Finally, Angelucci (2005) conducted an analysis of the Spanish version of the GSS 1, administering it to a sample of 320 Venezuelan university students. The results revealed the presence of two underlying components in the scale: the first with an eigenvalue of 2.60, explaining 13 % of the total variance, and the second with an eigenvalue of 1.75, accounting for 8.7 % of the variance. This result is partly consistent with the findings of Gudjonsson (1984), since although the factor structure obtained differed in terms of the grouping of items measuring suggestibility, it was similar concerning the true questions. Given these findings, the author concluded that the identified factors exhibited low internal consistency ($\alpha = .62$ and $\alpha = .47$, respectively), highlighting the need for further research.

Regarding Validity Evidence Based on Relations with External Variables (EBE) for the GSS 1 and GSS 2, in the study by Leue and Bertram (2024), a path analysis was conducted to examine a theoretical

model linking working memory capacity (assessed via the I-S-T 2000 R), immediate recall of story details from the GSS-1, and the GSS-1 suggestibility subscales (Yield 1, Yield 2, and Shift). Model fit was excellent (RMSEA = .00, CFI = 1.00, SRMR = .05), indicating strong alignment with the observed data. Working memory was positively associated with the number of correctly recalled details ($\beta = 0.31$, $p < .01$). However, recall performance did not significantly predict Yield 1 ($\beta = 0.11$, $p = .42$), Yield 2 ($\beta = -0.11$, $p = .45$), or Shift ($\beta = 0.22$, $p = .08$). These findings suggest that, while working memory supports initial story recall, suggestibility is an independent construct not directly determined by recall ability.

This systematic review identified only three studies that addressed validity evidence based on relations with external variables. One of them, Merkelbach et al. (1998), examined the predictive validity of Yield scores from the GSS 1, using a sample of 53 undergraduate students. The mean Yield score for this sample was 4.0 ($SD = 1.2$), while the mean score on suggestive questions related to the critical slide was 1.4 ($SD = 1.4$). The Pearson correlation between participants' Yield scores and the extent to which they yielded to key questions was small but significant ($r = 0.22$, $p < .05$).

In the study conducted by Bianco and Curci (2015), the GSS 1 and GSS 2 were administered alongside cognitive, personality, and affective measures (such as coping strategies and anxiety) to assess discriminant (divergent) validity. The results suggest that GSS 1 and GSS 2 effectively differentiate between IS and other cognitive measures. Moreover, the lack of a strong association between GSS scores and personality traits suggests that, when measured in an Italian sample, IS may represent a distinct construct from personality. Additionally, no significant correlations were found between suggestibility and anxiety, contradicting Gudjonsson's (1984) initial hypothesis while aligning with other studies (Haraldsson, 1985; Pires et al., 2013; Polczyk, 2005). Similarly, no correlations were observed between suggestibility and coping orientation.

Finally, the second part of the research by Leue and Bertram (2024), examined relationships among the GSS-1, working memory ability (assessed via the I-S-T 2000 R), and the Conflict-Monitoring Questionnaire (CMQ-28). The total number of correctly recalled details from the GSS-1 story was positively correlated with working memory ($r = .31$, $p < .01$), indicating that recall performance and working memory capacity may reflect convergent constructs. However, the GSS-1 subscales (Yield 1, Yield 2, and Shift) were not significantly correlated with working memory or CMQ-28 scales ($rs(101) = .09-.17$), suggesting that these subscales measure constructs distinct from memory capacity, consistent with the theoretical model of the GSS. Regarding reliability evidence (EC) based on internal consistency, the most commonly used measure in the analyzed studies was Cronbach's alpha coefficient. Among the reviewed works, only one study employed the Kuder-Richardson coefficient (K-R 20) (Bianco & Curci, 2015), while another conducted an omega reliability analysis (McDonald's ω) (Gignac & Powell, 2009).

Regarding the objective of this systematic review of analyzing the reliability of the GSS 1 and 2 scales, one of the earliest published alpha calculations for the GSS was conducted by Gudjonsson (1984), who reported Cronbach's alpha coefficients of .77 for the 15 Yield items and .67 for the 15 Shift items in a sample of $N = 195$. Given the lower reliability of the Shift factor, Singh and Gudjonsson (1987) modified the scoring method by incorporating five previously excluded affirmative items, which increased the Shift alpha to .71. However, it is well-established in the literature that increasing the number of items can artificially inflate alpha values, without necessarily improving the instrument's precision (Damásio & Borsa, 2017). Therefore, this adjustment does not guarantee greater measurement accuracy.

Bianco and Curci (2015) reported Cronbach's alpha values for the subscales Immediate Recall (.95) and Delayed Recall (.93), both of which are well above the commonly accepted threshold of .70. According to Tavakol and Dennick (2011), very high alpha values (above .90) may indicate redundancy among some items. In this study, the Kuder-Richardson coefficient was also calculated for the subscales: Yield 1 (.66), Yield 2 (.71), Shift (.58), and Total Suggestibility (.68). These values are considered questionable, as they fall below the average of the coefficients reported in the other studies.

The study by Gignac and Powell (2009) was the only one to present McDonald's Omega values. The Yield 1 subscale obtained a McDonald's Omega (ω) = .83, suggesting a respectable level of internal consistency reliability. In contrast, the Shift factor's McDonald's Omega was below the recommended threshold ($\omega = .60$), indicating an inadequate level of internal consistency reliability. As a result, the authors advise against using the Shift subscale for measuring suggestibility, and caution is recommended when utilizing the GSS in assessments.

In the present study, only Cronbach's Alpha was meta-analyzed due to its higher occurrence in the identified studies, as it is the most commonly reported index in the literature and still the most widely used for verifying the internal consistency of psychological instruments. However, it is suggested that in future studies evaluating the GSS, other internal consistency indicators (such as McDonald's Omega and composite reliability) should be considered.

All Cronbach's alpha values were meta-analyzed and, overall, indicated moderate (.75-.79) to fair (.70-.74) reliability according to established criteria (excellent $\geq .85$; good .80-.84; moderate .75-.79; fair .70-.74 (Ponterotto & Ruckdeschel, 2007; Santos et al., 2020). These results provide preliminary evidence of reliability, although this evidence is accompanied by substantial heterogeneity across studies. High heterogeneity (elevated I^2) among alpha values suggests considerable variability, likely reflecting methodological, population, or contextual differences. Subgroup comparisons, including across the two GSS versions, may help to explain part of this heterogeneity.

It is important to note that some analyses revealed very high levels of heterogeneity ($I^2 > 75\%$), indicating that the variability in reliability estimates across studies is largely attributable to true between-study differences rather than sampling error. Under such conditions, pooled estimates should be interpreted with caution, as they may not represent a single underlying population parameter but rather a distribution of context-dependent values. This level of heterogeneity limits the generalizability and precision of the meta-analytic findings and underscores the importance of considering study-level and sample-level characteristics as potential sources of variability. Accordingly, the results should be viewed as indicative of variability in the psychometric performance of the GSS across contexts, rather than as definitive evidence of stable reliability properties (Borenstein et al., 2009; Higgins et al., 2003).

Regarding test-retest reliability, Merckelbach et al. (1998) reported a correlation of .55 ($p < .01$), considered inadequate, whereas Gudjonsson (1987) found a strong correlation between GSS-1 and GSS-2 ($r = .92, p < .001$), indicating that IS may represent a stable behavioral characteristic over time.

Another approach to assessing the reliability of psychometric instruments is the split-half method. Hein et al. (2021) reported Spearman-Brown corrected split-half coefficients of Yield 1 = .45, Yield 2 = .71, and Shift = .68, with the minimum recommended threshold being $\geq .70$ (Zanon & Hauck-Filho, 2015). Yield 1 was slightly below this benchmark, whereas Yield 2 and Shift met or exceeded it. Factor reliability estimates (Squared Factor Score Determinacy, Hancock's H) were Yield 1: $H = .59$, Yield 2: $H = .71$, and Shift: $H = .71$, indicating similar patterns relative to the recommended range of $H = .70-.80$. In a separate study, Leue and Bertram (2024) reported a Spearman-Brown corrected split-half reliability of .72 and odd-even split reliability of .88 for the immediate recall scales (40 details) of the GSS-1, supporting adequate internal consistency for this measure.

Finally, inter-rater reliability for the GSS was consistently high across studies. Wachi et al. (2019b) reported $r = .99$, with Kappa scores for Yield 1, Yield 2, and Shift of .97, .99, and .97, respectively. For delayed recall on day two, the first coder scored all samples ($n = 218$) and the second coded 30 % ($n = 66$), yielding $r = .99$. Richardson and Smith (1993) assessed $n = 57$ English participants, reporting Yield 1 $r = .98$, Yield 2 $r = .99$, Shift $r = .95$, and Total Suggestibility $r = .99$, with only 1.4 % disagreement across 2,280 individual item ratings. Clare et al. (1994) examined $n = 101$ participants, finding intraclass correlation coefficients of .97 (immediate recall), .95 (delayed recall), and .99 for Yield 1, Yield 2, Shift, and Total Suggestibility (all $p < .001$). These findings demonstrate excellent agreement and strong inter-rater reliability across evaluators. The influence of the interviewer on suggestibility scores has been examined in several studies (Bain et al., 2004; Baxter et al., 2003, 2006). Clare et al. (1994), noted that when interviewers were trained by the scale's creator, Gudjonsson, gaps remained for practitioners relying solely on the manual or validation studies, highlighting the importance of comprehensive training to ensure accurate administration and scoring.

Practical implications

These studies underscore important implications for legal practice, as interviewees—whether witnesses or suspects—may be led to recall errors by interviewers or other external influences. Such effects can be mitigated by encouraging free recall and emphasizing focus on the original story while minimizing social pressure (Clare et al., 1994; Richardson & Smith, 1993; Özcan & Ziyalar, 2023). Further research aimed at improving applicability in forensic contexts is warranted. Evidence from online administration (Hein et al., 2021; Wachi et al., 2019a; Wachi et al., 2019b) suggests that digital versions of the GSS may offer a viable alternative to in-person administration. The online format could

reduce interviewer-related bias and standardize feedback delivery, improving procedural consistency. However, further research aimed at enhancing the applicability of the GSS in forensic contexts is warranted, particularly regarding the Shift dimension, which showed low internal consistency in some studies. Therefore, caution is recommended when interpreting Shift scores, and the GSS should not be used as the sole basis for decision-making in forensic assessments or investigative processes.

Study limitations

This systematic review and meta-analysis employed rigorous, structured methods to minimize bias, consistent with best practices for evidence synthesis (Egger et al., 1997; Higgins et al., 2019). Nonetheless, some limitations are acknowledged. A key limitation in this study is the high heterogeneity observed in the meta-analysis, with alpha values ranging from .43 to .67. Such variability may reflect differences in the construct measured across studies or substantial variation in study conditions, populations, or methodologies. Inter-rater agreement was quantified using percent agreement, which reached 74.2 % at the title/abstract screening stage. It is acknowledged that no formal kappa coefficient was calculated, which constitutes a methodological limitation of this study.

Conclusion

The findings of this systematic review and meta-analysis suggest that the Gudjonsson Suggestibility Scale (GSS) demonstrates moderate evidence of reliability, although this evidence is characterized by substantial heterogeneity across studies, indicating that reliability estimates may vary considerably depending on sample and contextual factors. In particular, the Shift factor consistently exhibited lower internal consistency, raising concerns regarding its precision and stability. Regarding validity, the available evidence remains limited, especially in relation to external variables, with a small number of studies providing such data. Therefore, current evidence is insufficient to support robust conclusions about the construct validity of the GSS across different contexts.

Overall, these findings indicate that, while the GSS may be useful for research purposes, its psychometric properties should be interpreted with caution. Future studies are needed to strengthen the evidence base, particularly by examining validity in relation to external criteria and by exploring sources of heterogeneity in reliability estimates.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf
- Angelucci, L. (2005). Escala de sugestionabilidad interrogativa de Gudjonsson: Confiabilidad, análisis de ítems y estructura factorial. *Interamerican Journal of Psychology*, 39(1), 79–88.
- Bain, S. A., Baxter, J. S., & Fellowes, V. (2004). Interacting influences on interrogative suggestibility. *Legal and Criminological Psychology*, 9(2), 239–252. <https://doi.org/10.1348/1355325041719419>
- Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence-Based Mental Health*, 22(4), 153–160. <https://doi.org/10.1136/ebmental-2019-300117>
- Baxter, J. S., Boon, J. C. W., & Marley, C. (2006). Interrogative pressure and responses to minimally leading questions. *Personality and Individual Differences*, 40(1), 87–98. <https://doi.org/10.1016/j.paid.2005.06.017>
- Baxter, J. S., Jackson, M., & Bain, S. A. (2003). Interrogative suggestibility: Interactions between interviewees' self-esteem and interviewer style. *Personality and Individual Differences*, 35(6), 1285–1292. [https://doi.org/10.1016/S0191-8869\(02\)00349-5](https://doi.org/10.1016/S0191-8869(02)00349-5)
- Bianco, A., & Curci, A. (2015). Measuring interrogative suggestibility with the Italian version of the Gudjonsson Suggestibility Scales (GSS): Factor structure and discriminant validity. *Personality and Individual Differences*, 82, 258–265. <https://doi.org/10.1016/j.paid.2015.03.031>
- Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, 27(4), 335–340. <https://doi.org/10.3102/10769986027004335>

- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. <https://doi.org/10.1002/9780470743386>
- Bottenberg, E. H., & Wehner, E. G. (1972). Suggestibilität. II. Einige persönlichkeits- und leistungsdiagnostische Korrelate des Würzburger Suggestibilitäts-Tests (WST). *Praxis der Kinderpsychologie und Kinderpsychiatrie*, 21(8), 282-288.
- Brink, Y., & Louw, Q. A. (2012). Clinical instruments: Reliability and validity critical appraisal. *Journal of Evaluation in Clinical Practice*, 18(6), 1126-1132. <https://doi.org/10.1111/j.1365-2753.2011.01707.x>
- Chien, P. F. W., Khan, K. S., & Siassakos, D. (2012). Registration of systematic reviews: PROSPERO. *BJOG: An International Journal of Obstetrics & Gynaecology*, 119(8), 903-905. <https://doi.org/10.1111/j.1471-0528.2011.03242.x>
- Clare, I. C. H., Gudjonsson, G. H., Rutter, S. C., & Cross, P. (1994). The inter-rater reliability of the Gudjonsson Suggestibility Scale (Form 2). *British Journal of Clinical Psychology*, 33(3), 357-365. <https://doi.org/10.1111/j.2044-8260.1994.tb01132.x>
- Cruz, M. P., & Pinho, M. S. (2014). Sugestionabilidade em pessoas idosas: Um estudo com a Escala de Sugestionabilidade de Gudjonsson (GSS 1). *Análise Psicológica*, 32(2), 199-213. <https://doi.org/10.14417/ap.613>
- Cunha, A. Q., Albuquerque, P. B., & Freire, T. (2007). Sugestionabilidade em crianças: Definição de conceitos e análises de variáveis cognitivas. *Psychologica*, 46, 125-141. <https://hdl.handle.net/1822/37662>
- Damásio, B. F., & Borsa, J. C. (Eds.). (2017). *Manual de desenvolvimento de instrumentos psicológicos*. Vetor.
- Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3), 399-412. <https://doi.org/10.1111/bjop.12046>
- Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629-634. <https://doi.org/10.1136/bmj.315.7109.629>
- Ferreira, B. G. (2018). *Psicologia do testemunho: Nos trilhos da mentira em busca da verdade* [Dissertação de mestrado, Universidade Lusófona de Humanidades e Tecnologias]. <https://recil.ulusofona.pt/server/api/core/bitstreams/198e289a-1f2c-4228-a945-939df27b8df7/content>
- Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63(3), 665-694. <https://doi.org/10.1348/000711010X502733>
- Gava, L. L., Pelisoli, C., & Dell'Aglio, D. D. (2013). A perícia psicológica em casos de suspeita de abuso sexual infanto-juvenil. *Avaliação Psicológica*, 12(2), 137-145.
- Gignac, G. E., & Powell, M. B. (2009). A psychometric evaluation of the Gudjonsson Suggestibility Scales: Problems associated with measuring suggestibility as a difference score composite. *Personality and Individual Differences*, 46(1), 88-93. <https://doi.org/10.1016/j.paid.2008.09.007>
- Gorassini, D. R., Harris, J. A., Diamond, A., & Flynn-Dastoor, E. (2006). Computer assessment of interrogative suggestibility. *Personality and Individual Differences*, 40(3), 569-577. <https://doi.org/10.1016/j.paid.2005.08.003>
- Gudjonsson, G. H. (1984). A new scale of interrogative suggestibility. *Personality and Individual Differences*, 5(3), 303-314. [https://doi.org/10.1016/0191-8869\(84\)90069-2](https://doi.org/10.1016/0191-8869(84)90069-2)
- Gudjonsson, G. H. (1987). A parallel form of the Gudjonsson Suggestibility Scale. *British Journal of Clinical Psychology*, 26(3), 215-221. <https://doi.org/10.1111/j.2044-8260.1987.tb01348.x>
- Gudjonsson, G. H. (1992). Interrogative suggestibility: Factor analysis of the Gudjonsson Suggestibility Scale (GSS 2). *Personality and Individual Differences*, 13(4), 479-481. [https://doi.org/10.1016/0191-8869\(92\)90077-3](https://doi.org/10.1016/0191-8869(92)90077-3)
- Gudjonsson, G. H. (1997). *The Gudjonsson suggestibility scales manual*. Psychology Press.
- Gudjonsson, G. H. (2003). *The psychology of interrogations and confessions: A handbook*. John Wiley & Sons.
- Gudjonsson, G. H., & Clark, N. K. (1986). Suggestibility in police interrogation: A social psychological model. *Social Behaviour*, 1(2), 83-104.

- Gudjonsson, G. H., & Singh, K. K. (1984). Interrogative suggestibility and delinquent boys: An empirical validation study. *Personality and Individual Differences*, 5(4), 425-430. [https://doi.org/10.1016/0191-8869\(84\)90007-2](https://doi.org/10.1016/0191-8869(84)90007-2)
- Gudjonsson, G. H., Rutter, S. C., & Clare, I. C. H. (1995). The relationship between suggestibility and anxiety among suspects detained at police stations. *Psychological Medicine*, 25(4), 875-883. <https://doi.org/10.1017/S0033291700035133>
- Gudjonsson, G. H., Vagni, M., Maiorano, T., & Pajardi, D. (2016). Age and memory related changes in children's immediate and delayed suggestibility using the Gudjonsson suggestibility scale. *Personality and Individual Differences*, 102, 25-29. <https://doi.org/10.1016/j.paid.2016.06.029>
- Gudjonsson, G. H., Vagni, M., Maiorano, T., & Pajardi, D. (2020). The relationship between trauma symptoms and immediate and delayed suggestibility in children who have been sexually abused. *Journal of Investigative Psychology and Offender Profiling*, 17(2), 155-168. <https://doi.org/10.1002/jip.1554>
- Hair, J. F., Jr., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2021). *A primer on partial least squares structural equation modeling (PLS-SEM)* (2nd ed.). Sage.
- Haraldsson, E. (1985). Interrogative suggestibility and its relationship with personality, perceptual defensiveness and extraordinary beliefs. *Personality and Individual Differences*, 6(6), 765-767. [https://doi.org/10.1016/0191-8869\(85\)90087-X](https://doi.org/10.1016/0191-8869(85)90087-X)
- Hein, F., Scheuble, V., Beauducel, A., & Leue, A. (2021). Psychometric properties of a German online version of the Gudjonsson Suggestibility Scale 1. *Frontiers in Psychology*, 12, 718805. <https://doi.org/10.3389/fpsyg.2021.718805>
- Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Wiley-Blackwell.
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414), 557-560. <https://doi.org/10.1136/bmj.327.7414.557>
- Hupe, M. (2019). EndNote X9. *Journal of Electronic Resources in Medical Libraries*, 16(3-4), 117-119. <https://doi.org/10.1080/15424065.2019.1691963>
- Kotov, R. I., Bellman, S. B., & Watson, D. B. (2004). *Multidimensional Iowa Suggestibility Scale (MISS)*. Stony Brook University.
- Leue, A., & Bertram, F. (2024). Working memory ability, suggestibility and conflict monitoring: On psychometrics and the nomological network. *Personality and Individual Differences*, 227, 112816. <https://doi.org/10.1016/j.paid.2024.112816>
- Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the misinformation effect. *Learning & Memory*, 12(4), 361-366. <https://doi.org/10.1101/lm.94705>
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412-433. <https://doi.org/10.1037/met0000144>
- Merckelbach, H., Muris, P., Wessel, I., & Van Koppen, P. J. (1998). The Gudjonsson Suggestibility Scale (GSS): Further data on its reliability, validity, and metacognition correlates. *Social Behavior and Personality: An International Journal*, 26(2), 203-210. <https://doi.org/10.2224/sbp.1998.26.2.203>
- Milne, R., Clare, I. C. H., & Bull, R. (2002). Interrogative suggestibility among witnesses with mild intellectual disabilities: The use of an adaptation of the GSS. *Journal of Applied Research in Intellectual Disabilities*, 15(1), 8-17. <https://doi.org/10.1046/j.1360-2322.2001.00096.x>
- Mokkink, L. B., Prinsen, C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C., & Terwee, C. B. (2018). *COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs): User manual*. COSMIN.
- Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan – a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1), 210. <https://doi.org/10.1186/s13643-016-0384-4>
- Özcan, E. I., & Ziyalar, N. (2023). Gudjonsson Suggestibility Scale-2 Turkish reliability and validity study. *Adli Tıp Bülteni*, 28(1), 6-14. <https://doi.org/10.17986/blm.1607>
- Parris, B. A. (2016). The prefrontal cortex and suggestion: Hypnosis vs. placebo effects. *Frontiers in Psychology*, 7, 415. <https://doi.org/10.3389/fpsyg.2016.00415>

- Pires, R., Silva, D. R., & Ferreira, A. S. (2013). Portuguese adaptation of the Gudjonsson Suggestibility Scales (GSS1 and GSS2): Empirical findings. *Personality and Individual Differences*, 54(2), 251-255. <https://doi.org/10.1016/j.paid.2012.09.008>
- Polczyk, R. (2005). Interrogative suggestibility: Cross-cultural stability of psychometric and correlational properties of the Gudjonsson Suggestibility Scales. *Personality and Individual Differences*, 38(1), 177-186. <https://doi.org/10.1016/j.paid.2004.03.018>
- Polczyk, R. (2016). Factor structure of suggestibility revisited: New evidence for direct and indirect suggestibility. *Current Issues in Personality Psychology*, 4(2), 97-104. <https://doi.org/10.5114/cipp.2016.60249>
- Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills*, 105(3), 997-1014. <https://doi.org/10.2466/pms.105.3.997-1014>
- Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, 27(5), 1147-1157. <https://doi.org/10.1007/s11136-018-1798-3>
- Raver, A., Lindholm, T., Gustafsson, P. U., & Alm, C. (2023). Memory accuracy, suggestibility and credibility in investigative interviews with native and non-native eyewitnesses. *Frontiers in Psychology*, 14. <https://doi.org/10.3389/fpsyg.2023.1240822>
- Richardson, G., & Smith, P. (1993). The inter-rater reliability of the Gudjonsson Suggestibility Scale. *Personality and Individual Differences*, 14(1), 251-253. [https://doi.org/10.1016/0191-8869\(93\)90197-B](https://doi.org/10.1016/0191-8869(93)90197-B)
- Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(4), 803-814. <https://doi.org/10.1037/0278-7393.21.4.803>
- Rovinski, S. L., & Cruz, R. M. (2009). *Psicologia jurídica: Perspectivas teóricas e processos de intervenção*. Vetor.
- Rowell, K., & Colloff, M. F. (2024). Critique of a measure of interrogative suggestibility for children: The Bonn test of statement suggestibility. *Applied Cognitive Psychology*, 38(2), 408-417. <https://doi.org/10.1002/acp.4145>
- Santos, G. M., Strathdee, S. A., El-Bassel, N., Patel, P., Subramanian, D., Horyniak, D., Cook, R. R., McCullagh, C., Marotta, P., Choksi, F., Kang, B., Allen, I., & Shoptaw, S. (2020). Psychometric properties of measures of substance use: A systematic review and meta-analysis of reliability, validity and diagnostic test accuracy. *BMC Medical Research Methodology*, 20(1), 1-22. <https://doi.org/10.1186/s12874-020-01019-8>
- Saraiva, M., & Albuquerque, P. B. (2015). Influência da idade, desejabilidade social e memória na sugestibilidade infantil. *Psicologia: Reflexão e Crítica*, 28, 356-364. <https://doi.org/10.1590/1678-7153.201528216>
- Shor, R. E., & Orne, E. C. (1962). *Harvard group scale of hypnotic susceptibility: Form A*. Consulting Psychologists Press.
- Singh, K. K., & Gudjonsson, G. H. (1987). The internal consistency of the "shift" factor on the Gudjonsson Suggestibility Scale. *Personality and Individual Differences*, 8(2), 265-266. [https://doi.org/10.1016/0191-8869\(87\)90183-8](https://doi.org/10.1016/0191-8869(87)90183-8)
- Singh, K. K., & Gudjonsson, G. H. (1992). Interrogative suggestibility among adolescent boys and its relationship with intelligence, memory, and cognitive set. *Journal of Adolescence*, 15(2), 155-161. [https://doi.org/10.1016/0140-1971\(92\)90023-C](https://doi.org/10.1016/0140-1971(92)90023-C)
- Souza, R. L. S., & Cruz, S. T. M. (2023). Perícia psicológica de vítimas e testemunhas de violência. In D. Lino (Org.), *Psicologia investigativa: Teoria e prática* (pp. 55-70). Juruá Editora.
- Stein, L. M., Pergher, G. K., & Feix, L. F. (2009). *Desafios da oitiva de crianças no âmbito forense*. Secretaria Especial dos Direitos Humanos da Presidência da República/Subsecretaria de Promoção dos Direitos da Criança e do Adolescente, Programa Nacional de Enfrentamento da Violência Sexual contra Crianças e Adolescentes, Childhood Brasil.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55. <https://doi.org/10.5116/ijme.4dfb.8dfd>

- Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, 60(1), 34-42. <https://doi.org/10.1016/j.jclinepi.2006.03.012>
- Wachi, T., Watanabe, K., Yokota, K., Otsuka, Y., & Hirama, K. (2019a). Comparison between Japanese online and standard administrations of the Gudjonsson Suggestibility Scale 2 and effects of post-warning. *Legal and Criminological Psychology*, 24(1), 71-86. <https://doi.org/10.1111/lcrp.12147>
- Wachi, T., Watanabe, K., Yokota, K., Otsuka, Y., & Hirama, K. (2019b). The immediate and delayed suggestibility examined by the online version of the Gudjonsson Suggestibility Scale 2. *Personality and Individual Differences*, 146, 20-25. <https://doi.org/10.1016/j.paid.2019.03.040>
- Weitzenhoffer, A. M., & Hilgard, E. R. (1959). *Stanford Hypnotic Susceptibility Scales*. Consulting Psychologists Press.
- Young, K., Powell, M. B., & Dudgeon, P. (2003). Individual differences in children's suggestibility: A comparison between intellectually disabled and mainstream samples. *Personality and Individual Differences*, 35(1), 31-49. [https://doi.org/10.1016/S0191-8869\(02\)00138-1](https://doi.org/10.1016/S0191-8869(02)00138-1)
- Zanon, C., & Hauck-Filho, N. (2015). Fidedignidade. In C. S. Hutz, D. R. Bandeira, & C. M. Trentini (Eds.), *Psicometria* (pp. 85-98). Artmed.

Authors' contribution (CRediT Taxonomy): 1. Conceptualization; 2. Data curation; 3. Formal Analysis; 4. Funding acquisition; 5. Investigation; 6. Methodology; 7. Project administration; 8. Resources; 9. Software; 10. Supervision; 11. Validation; 12. Visualization; 13. Writing: original draft; 14. Writing: review & editing.

R. L. S. S. has contributed in 1, 2, 3, 5, 6, 8, 11, 12, 13, 14; M. P. G. in 1, 2, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14; F. N. dos S. in 2, 3, 5, 6, 13; J. C. A. in 6, 10, 11, 12, 14; J. L. C. J. in 2, 3, 5, 6, 9, 13, 14; R. de F.-D. in 1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 13, 14.

Scientific editor in charge: Dra. Cecilia Cracco.

Appendix A

Table A1

Search Strategies by Database (February 2025)

Database	Platform	Search Query	Records (n)
Scopus	Elsevier	TITLE-ABS ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales")	181
Academic Search Premier	EBSCOhost	TI ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales") OR AB ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales")	106
Embase	Elsevier	('Gudjonsson Suggestibility Scale':ab,ti OR 'Gudjonsson Suggestibility Scales':ab,ti)	84
PubMed	NLM	("Gudjonsson Suggestibility Scale"[Title/Abstract] OR "Gudjonsson Suggestibility Scales"[Title/Abstract])	44
MEDLINE	EBSCOhost	TI ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales") OR AB ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales")	42
CINAHL	EBSCOhost	TI ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales") OR AB ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales")	16
APA PsycArticles (APA Journals)	APA	TI ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales") OR AB ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales")	14
SPORTDiscus	EBSCOhost	TI ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales") OR AB ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales")	3
LILACS	BVS/BIREME	(ti:("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales") OR ab:("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales"))	1
PsycINFO	APA	TI ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales") OR AB ("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales")	0
PePSIC	BVS/BIREME	(ti:("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales") OR ab:("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales"))	0
SciELO	BVS/BIREME	(ti:("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales") OR ab:("Gudjonsson Suggestibility Scale" OR "Gudjonsson Suggestibility Scales"))	0

Table A2

List of excluded studies

Study title	Year	Authors
Interrogative suggestibility among adolescent boys and its relationship with intelligence, memory, and cognitive set	1992	K. K. Singh, & G. H. Gudjonsson
The Portuguese adaptation of the Gudjonsson Suggestibility Scale (GSS1) in a sample of inmates	2014	R. Pires, D. R. Silva, & A. Sousa Ferreira
Investigation of the responses of 'fakers' and 'non-fakers' on the Gudjonsson Suggestibility Scale (GSS)	1986	K. Smith, & G. H. Gudjonsson
Interrogative suggestibility in the elderly	2020	S. Biondi, C. Mazza, E. Ricci, G. Orru, A. Di Domenico, M. Monaro, & P. Roma
Memory accuracy, suggestibility and credibility in investigative interviews with native and non-native eyewitnesses	2023	A. Raver, T. Lindholm, P. U. Gustafsson, & C. Alm
Further data on interrogative suggestibility and compliance scores following instructed malingering	2010	I. Hansen, T. Smeets, & M. Jelicic
How well does the Gudjonsson Suggestibility Scale for Children, version 2 predict the recall of false details among children with and without intellectual disabilities?	2007	K. L. Miles, M. B. Powell, G. E. Gignac & D. M. Thomson
Computer assessment of interrogative suggestibility	2006	D. R. Gorassini, J. Aitken Harris, A. Diamond, & E. Flynn-Dastoor
Interrogative suggestibility in a US context: Some preliminary data on normal subjects	2004	R. Pollard, B. Trowbridge, P. D. Slade, A. P. Streissguth, A. Laktonen, B. D. Townes
Comparison between Gudjonsson Suggestibility Scale 2 (GSS2) and Bonn Test of Statement Suggestibility (BTSS) in measuring children's interrogative suggestibility	2011	P. Romaa, U. Sabatello, G. Verrastrob, & S. Ferracutia
Minimizing extraneous, interviewer-based interrogative suggestibility	2004	J. C.W. Boon & J. S. Baxter
Opposing determinants of compliance and interrogative suggestibility	2013	J. S. Baxter, S. A. Bain, A. Pringlec, H. Fowlerc, & T. Tafilía
Patterns of malingering and compliance in measures of interrogative suggestibility	2006	R. Woolston, S. A. Bain, & J. S. Baxter
Sugestionabilidade em pessoa idosa: um estudo com a escala GSS1	2014	M. Pedreiras Cruz, & M. S. Pinho
Interrogative suggestibility: The role of interviewer behaviour	2000	S. A. Bain & J. S. Baxter
Shortened versions of the Gudjonsson Suggestibility Scale meet the standards	2009	T. Smeets, J. Leppink, M. Jelicic, & H. Merckelbach
Interrogative suggestibility: does the setting where subjects are tested make a difference to the scores on the GSS 1 and GSS 2?	1995	G. H. Gudjonsson
Interrogative suggestibility and its relationship with personality, perceptual defensiveness and extraordinary beliefs	1985	E. Haraldsson
Portuguese adaptation of the Gudjonsson Suggestibility Scales (GSS1 and GSS2): Empirical finding	2013	R. Pires, D. R. Silva, & A. Sousa Ferreira
Gudjonsson Suggestibility Scales	1997	G. H. Gudjonsson
The potential conflict between clinical and judicial decision-making heuristics	1999	E. Rassin, & H. Merckelbach
Interrogative suggestibility among witnesses with mild intellectual disabilities: the use of an adaptation of the GSS	2002	R. Milne; I. C. H. Clare, & R. Bul
The response alternatives of suggestible and non-suggestible adolescent offenders	1998	G. R. Richardson, T. P. Kelly, & A. Bryce
Interrogative suggestibility: Empirical findings	2003	G. H. Gudjonsson
Suggestibility and individual differences: Psychosocial and memory measures	2013	A. M, Ridley, & G. H. Gudjonsson
The role of interrogative suggestibility in suspect interview and eyewitness testimony: Applications of the Gudjonsson Suggestibility Scales	2016	A. Bianco, & A. Curci