Guidelines for the Assessment of Efficacy of Clinical Hypnosis Applications

Authors:

Zoltan Kekecs[1, 2]; Donald Moss[3], Gary Elkins[4] Giuseppe De Benedittis[5], Olafur S. Palsson[6], Philip D. Shenefelt[7], B. Devin Terhune[8] Katalin Varga[2], Peter J. Whorwell[9]

Affiliations:

[1]Department of Psychology, Lund University

[2]Institute of Psychology, ELTE Eötvös Loránd University

[2]Department of Psychology, Lund University

[3]College of Integrative Medicine and Health Sciences, Saybrook University

[4]Department of Psychology and Neuroscience, Baylor University

[5]Department of Neurosurgery, University of Milano

[6]Department of Medicine, University of North Carolina at Chapel Hill

[7] Department of Dermatology and Cutaneous Surgery, University of South Florida, Tampa...

[8]Department of Psychology, Goldsmiths, University of London

[9]Neurogastroenterology Unit, Wythenshawe Hospital

This is a manuscript currently undergoing peer review in a scholarly journal.

Corresponding author:

Zoltan Kekecs

Department of Psychology, Lund University

Institute of Psychology, ELTE Eötvös Loránd University;

email: zoltan.kekecs@psy.lu.se

Abstract

Research on the efficacy of hypnosis applications continues to grow, but there remain major gaps between the science and clinical practice. One challenge has been a lack of consensus on what applications of hypnosis are efficacious based on research evidence. In 2018, six major hypnosis organizations collaborated to form a Task Force for Establishing Efficacy Standards for Clinical Hypnosis. This paper describes a Guideline for the Assessment of Efficacy of Clinical Hypnosis Applications developed by the Task Force which makes ten specific recommendations. The guideline is intended to be a tool for those who want to assess the quality of existing evidence on the efficacy of clinical hypnosis for any particular indication. The paper also discusses methodological issues in the interpretation and implementation of these guidelines. Future papers will report on the other products of the Hypnosis Efficacy Task Force, such as best practice recommendations for outcomes research in hypnosis, and an international survey of researchers and clinicians on current practice and attitudes about hypnosis.

Keywords: clinical efficacy, research best practices, GRADE system, hypnosis

Guidelines for the Assessment of Efficacy of Clinical Hypnosis Applications

There is a wealth of research on clinical applications of hypnosis. Likewise, clinical hypnosis is used in the treatment of a multitude of disorders and illnesses by clinicians worldwide. However, there is a disconnect between the scientific literature and much of clinical practice (Jensen et al., 2017). Many of the specific applications of this treatment modality, even some of the ones that are widely used in clinical practice, have still not been investigated in research and are therefore not supported by scientific evidence. One of the reasons for this disconnect is that so far there have been no widely accepted standards for establishing the efficacy of clinical hypnosis interventions. Although double-blind controlled trials provide guardrails that reduce straying far from validity, they are not feasible for hypnosis trials. To address this issue, the Task Force for Establishing Efficacy Standards for Clinical Hypnosis (from hereon, the Hypnosis Efficacy Task Force) was assembled. In this paper we present a list of recommendations for researchers and clinicians who plan to assess the efficacy of clinical applications of hypnosis in the treatment of medical and mental health disorders and symptoms.

Healthcare providers, insurers, clinicians, and patients are looking for clear, evidence-based recommendations about which therapies to use. The field of clinical hypnosis is now at a point, after nearly a century of formal scientific hypnosis research, where hundreds of research trials and case studies investigating the effectiveness of hypnosis-based interventions for the treatment of various symptoms and conditions have been published (for recent reviews, see e.g., Madden et al., 2016; Carlson, et al., 2018; Kendrick et al., 2016; Fisch et al., 2017; Catsaros & Wendland, 2020). Thus, it is reasonable to expect researchers to be able to formulate evidence-based recommendations about clinical applications of hypnosis. Such recommendations, based

on the systematic evaluation of accumulated evidence, are integral for making decisions regarding the adoption of effective interventions. The standards for assessing the efficacy of interventions have evolved considerably since the emergence of debate on evidence-based practice and empirically supported interventions in the 1990s (Evidence-Based Medicine Working Group, 1992; Sackett et al., 1996; Chambless & Hollon, 1998).

Several evidence grading systems have emerged over the past decades for issuing clinical recommendations in both psychotherapy research (see, e.g., Chambless & Ollendick, 2001) and clinical medicine (e.g. Atkins et al., 2004; OCEBM Working Group, 2011). Nevertheless, to date there has been no consensus on standards for clinical efficacy determination in the hypnosis field, thereby preventing the field from issuing a clear and unequivocal message about the efficacy of treatment applications. This shortcoming has potentially played a role in limiting the utilization of hypnosis as a treatment option in healthcare in society in general.

Perhaps one reason for the lack of take-up of the above-mentioned evaluation methods in our field might be that there are some field-specific considerations in clinical hypnosis research that are not clearly addressed in these evidence grading systems. For example, it is not clear from these systems whether and how to take into consideration the hypnotizability of participants in the trials, and which studies can be taken into consideration in the efficacy assessment, when there are so many different intervention variants. Furthermore, double-blind placebo-controlled designs are held as the gold standard in most of the previous rating systems for demonstrating efficacy. However, the central role of expectancy in hypnotic effects demonstrated in both laboratory and clinical trials (e.g., Lynn et al., 2008) makes it unclear whether these types of designs would convey the same information about efficacy as in medical research.

Thus, in order to facilitate the adoption of efficacious clinical hypnosis interventions in healthcare, there is an urgent need for a consensus-based system for issuing evidence-based clinical recommendations about applications of clinical hypnosis. The recognition of this need led to the formation of the Hypnosis Efficacy Task Force.

The Hypnosis Efficacy Task Force

In 2018, in recognition of the unmet need for efficacy standards in field of hypnosis, the Society for Clinical and Experimental Hypnosis (SCEH) initiated an organizational meeting on this issue at the triennial Congress of the International Society of Hypnosis (ISH) in Montréal, Canada. Shortly thereafter, six major hypnosis societies agreed to co-sponsor an international "Task Force for Establishing Efficacy Standards for Clinical Hypnosis". Co-sponsors included SCEH, the American Society of Clinical Hypnosis, the American Psychological Association Division 30, the Milton Erickson Foundation, the National Pediatric Hypnosis Training Institute, and the International Society of Hypnosis.

Zoltan Kekecs and Donald Moss agreed to convene and guide the Task Force, which was composed of nine selected researchers from Hungary, the US, the UK, and Italy who committed to participating in the Task Force deliberations. The participants are the authors of this paper: Giuseppe De Benedittis, Gary Elkins, Zoltan Kekecs, Donald Moss, Olafur S. Palsson, Philip D. Shenefelt, Devin B. Terhune, Katalin Varga, and Peter Whorwell. In addition, additional researchers agreed to serve as consultants to the Task Force: Mark Jensen, Elvira Lang, and David Patterson.

The Task Force defined and pursued three objectives: 1) to establish guidelines for the assessment of the efficacy of hypnosis applications, based on methodological criteria; 2) to develop recommendations for best practices in future outcomes research on clinical hypnosis; and 3) to conduct an international survey of clinicians, researchers, and students in the field of hypnosis, to provide the most comprehensive picture to date on current practices and views in this community.

This paper will introduce the Hypnosis Efficacy Task Force's Guidelines for the Assessment of Efficacy of Clinical Hypnosis Applications that resulted from the work on the first objective. These guidelines are not intended to serve as a stand-alone system for efficacy assessment. Instead, they serve as recommendations for applying already existing comprehensive efficacy rating systems to data in the field of clinical hypnosis. The sections that follow provide a detailed description of the guidelines, the methods through which they were derived, and where appropriate, some rationale on why a specific guideline was chosen.

Methods

The Guidelines listed below are based on discussions in a series of monthly meetings of the Hypnosis Efficacy Task Force between February and November 2019. In the first meetings, the Hypnosis Efficacy Task Force reviewed existing evidence rating and recommendations systems, such as the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system (Guvatt et al., 2008), the OCEBM: Levels of Evidence Table (OCEBM Working Group, 2011), and the evidence grading system of the Association for Applied Psychophysiology and Biofeedback (AAPB) and the International Society for Neuronal Regulation (ISNR) (LaVaque et al., 2002; Tan et al., 2016), as well as the APA Division 12

Empirically Supported Therapies (ESTs) rating system (Chambless & Hollon, 1998). Based on this initial review, we decided that the GRADE system provides a suitable framework for synthesizing evidence and formulating clinical recommendations.

However, we concluded that additional work was needed to make this system applicable in the field of clinical hypnosis to take into consideration important hypnosis-specific research features that can influence the assessment of effectiveness, risk of bias, and quality of evidence. After this consensus decision, an initial list of recommendations was drafted, which was then reviewed, amended, and supplemented by the Task Force in subsequent meetings. The final draft of the guidelines was sent to the external consultants -- Mark Jensen, Elvira Lang, and David Patterson -- who reviewed the list and recommended improvements. These suggested amendments were integrated into the recommendations at subsequent meetings. The final wording of the guidelines was reviewed and approved unanimously at the November 26, 2019 meeting by all members who contributed to formulating the guidelines, namely Giuseppe De Benedittis, Gary Elkins, Zoltan Kekecs, Donald Moss, Olafur Palsson, Philip Shenefelt, Devin B. Terhune, Katalin Varga, and Peter Whorwell.

Guidelines for the Assessment of Efficacy of Clinical Hypnosis Applications

In this section we list the Hypnosis Efficacy Task Force's Guidelines for the Assessment of Efficacy of Clinical Hypnosis Applications. The following recommendations are intended to guide researchers who want to assess the accumulated evidence - based on multiple studies - about the efficacy of certain applications of clinical and medical hypnosis. The guidelines contain recommendations about which methods are thought to be adequate for the assessment of

efficacy and quality of evidence and highlight certain important features that should be taken into consideration during an efficacy assessment.

1. Establishment of efficacy should be based on a sufficiently recent systematic review matching the highest quality standards, including multiple studies supporting the effectiveness of the treatment application. Whenever possible, the systematic review should be accompanied by a quantitative synthesis of the effect sizes (such as a meta-analysis) at the time of publication. The systematic review on which the efficacy assessment is based needs to be peer-reviewed (a peer reviewed journal article or book chapter are both eligible).
2. GRADE guidelines are endorsed by the Hypnosis Efficacy Task Force to assess efficacy (Guyatt et al., 2008a, 2008b, 2008c).
    a. Note: It is possible that there is or will be in the future a system other than the GRADE that is appropriate to assess efficacy. Thus, the use of other systems is not specifically excluded. Nevertheless, the system must be comparable in sophistication and reliability to the GRADE system and must account for all potential biases considered in the GRADE system.
3. The sample size, effect size (and associated confidence intervals), and clinical significance should be taken into consideration when evaluating efficacy. Thus, the systematic review(s) and meta-analysis(es), on the basis of which efficacy is determined, should highlight all of this information. Furthermore, where relevant, data from non-completers within research studies should also be taken into

consideration when assessing efficacy (for example data reported via intention-to-treat [ITT] analysis).

4. The assessment of hypnotizability is encouraged in clinical hypnosis studies (Jensen et al., 2017), since it can inform about the underlying mechanisms producing any therapeutic effects but is not required to establish the efficacy of a hypnosis-based treatment.
5. Blinding of the participants/patients and the interventionists to group allocation is aspirational but is not required to establish efficacy of a hypnosis-based treatment.
    a. Note: However, establishing that a hypnosis-based intervention has benefits over a well-matched placebo/sham control condition, or an already established active treatment condition, in a study where participants were blinded to group allocation can strengthen inferences regarding the specificity of the intervention.
6. Blinding of data collectors with respect to group allocation and/or hypnotizability level of the participant reduces the risk of experimenter biases (Barber, 1976). This should be taken into consideration in the risk of bias assessment and when determining the quality of the evidence (see also Holman et al., 2015).
7. Blinding of those responsible for the statistical analysis with respect to group allocation can decrease the risk of experimenter biases. (Automation of the analysis or exact pre-registration of the analysis plan can serve the same purpose). This should be taken into consideration in the risk of bias assessment and when determining the quality of the evidence.

8. The efficacy rating of hypnosis applications should be based on publications that meet the following criteria: 1) the intervention (or a component of a complex intervention) is labeled by the authors of the paper as “hypnosis” or one of its close synonyms (“hypnotic treatment”, “hypnotherapy”, etc.); and 2) the description of the intervention does not describe a process that expert reviewers would not categorize as hypnotic, under current consensus (for a consensus-based definition, see, e.g., Elkins et al., 2015).
    a. Note: It is not necessary that the intervention has been labelled as “hypnosis” to the participants of the study, but the labeling of the intervention to participants should be considered as a possible moderator in the meta-analysis, since labeling the intervention as hypnosis to participants has been found to increase effect size (Gandhi & Oakley, 2005).
9. In order to warrant the highest quality of evidence rating, the studies supporting the efficacy of the treatment should be conducted by at least two independent research groups, or at least one of the studies supporting efficacy should be a multi-center clinical trial.
10. For chronic or enduring conditions, efficacy needs to be demonstrated at a long follow-up assessment which is considered clinically appropriate for the given condition to warrant the highest quality of evidence rating. For many conditions, such as chronic pain, this would be six months or longer.

Discussion

The list of recommendations presented above is deliberately concise in order to ensure its practical usefulness for researchers. Below we discuss some of the considerations that went into formulating these guidelines and other topics relevant to fully understand them.

GRADE

The efficacy assessment guidelines put together by the Hypnosis Efficacy Task Force endorse the GRADE system for assessing the level of evidence for efficacy, and for formulating clinical recommendations. The reason for this choice was two-fold. On the one hand, this system seemed well-developed and comprehensive. One distinguishing feature of GRADE is that it includes a systematic review of the research studies assessing the effectiveness of the clinical application, and making a decision based on all studies found in the systematic review combined, while most other systems only require a certain number of studies showing efficacy for the efficacy rating. On the other hand, GRADE is currently the most accepted clinical recommendation system in medical research, with many high quality journals including it in their standard submission guidelines. Since clinical hypnosis has a great number of medical applications, it is an added advantage that recommendations made using the GRADE system would be easier to understand and seen as more credible for medical professionals and decision makers than those made using systems they are less familiar with, such as the Division 12 ESTs system primarily devised for psychotherapy applications. Tolin and colleagues (Tolin, McKay, Forman, Klonsky & Thombs, 2015) provide a good overview of the criticisms of the Division 12 ESTs system in its original form, and rationale for why the ESTs need to be updated in a way that they are based on a systematic review of the literature and on the GRADE system.

Conducting a GRADE review is time-consuming and needs to be planned prospectively before the systematic review is conducted. Thus, before conducting evaluation of the efficacy of a hypnosis application, researchers need to familiarize themselves with GRADE. This can be done by following instructions on the GRADE Working Group's website: https://www.gradeworkinggroup.org/, and by reading the main publications on the method (Balshem et al., 2011; Guyatt et al., 2008a; Guyatt et al., 2008b; Guyatt et al., 2008c; Guyatt et al., 2011a; Guyatt et al., 2011b). For a concise overview of the GRADE system, see https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/ by Siemieniuk and Guyatt (n.d.). The readers can also find good guidance about how Cochrane Reviews and GRADE recommendations can be integrated by the Cochrane GRADEing Methods Group (Schünemann, et al., 2019).

The following is a high-level summary of the GRADE system: The researchers conduct a systematic review of the research studies conducted on the clinical application of interest and, following specific instructions, produce two main outcomes: 1) they state the quality of evidence supporting the efficacy of the application, and 2) they issue a recommendation about the use of the intervention for treating the symptom or condition. The quality of the evidence is rated on a four-level scale ("very low," "low," "moderate," "high") depending on a number of factors such as study limitations, consistency and precision of results, directness of evidence, publication bias, and magnitude of the effect. The meaning of the different quality of evidence ratings are provided in Table 1, based on Siemieniuk and Guyatt (2021). In a review only including randomized controlled trials, the quality of evidence rating starts out at high but can be down-graded if there are concerns related to risk of bias in the studies, imprecision of effect estimates,

inconsistency of the findings among the reviewed studies, indirectness of evidence due to the studied populations being not directly relevant, or publication bias. The quality of evidence can be up-graded if the studies indicate a very large magnitude of effect, if there is evidence for dose-response in the studies, or when residual confounding is likely to decrease rather than increase the magnitude of effect (for more details, see Box 1. and Schünemann, Brozek, Guyatt & Oxman, 2013). A "high" quality of evidence rating is very rarely given due to the high standards this requires, and we do not anticipate that at present many hypnosis applications would receive this rating, but as new, high-quality research evidence is accumulated, more and more applications may reach this level.

In addition to the quality of evidence rating, a GRADE recommendation regarding an intervention can be either "strong recommendation" or "weak recommendation" to use the treatment, or the reviewers can issue a strong or weak recommendation against the use of the treatment. When determining the level of recommendation, the reviewers need to consider factors such as balance between desirable and undesirable effects, quality of evidence, values and preferences of patients, and costs of treatment. Table 2. provides more details about the influence of these factors on the strength of recommendation based on Guyatt et al. (2008a). For example, if it is clear that the benefits far outweigh the risks and virtually all informed patients would make the same choice, a "strong recommendation" would be issued. In contrast, if considering the evidence, most informed patients would choose to use the treatment, but a substantial number would not choose it for some reason, so patient values and preferences will play a crucial role in the final decision by the patient, a "weak recommendation" would be issued (Andrews et al., 2013). A "strong recommendation" may be issued even if the quality of

evidence is not “high”. Rather, a recommendation level will depend on the balance between the benefits and the costs and risks associated with the application. In addition, it should be noted that applications with a “weak recommendation” are still recommended. It is just that personal values and preferences tend to play a larger role in choosing the treatment compared to treatments with “strong recommendation”, which are basically “no-brainers”.

Despite our efforts we did not find another assessment system that would be comparable in sophistication and sensitivity to bias to the GRADE system, so currently this is the only system that is endorsed by the Task Force. If another system is used, it must be comparable in sophistication and reliability to the GRADE system and must account for all of the potential biases that are considered in the GRADE system.

Clinical significance

It is important to note that clinical recommendations are not primarily based on statistical significance. Demonstrating statistically significant evidence supporting the treatment effect is a necessary but not a sufficient condition for recommending the use of the treatment. Even a very small and clinically meaningless effect can be statistically significant depending on the size of the sample and the variability in the population. Thus, aside from statistical significance, the reviewers also have to consider the clinical significance of the treatment effects. Judging whether the treatment effect constitutes a clinically meaningful change requires specialized knowledge about the patient population, the illness or problem being treated, as well as the different measures used to assess the clinical outcomes and how these compare to each other. In some cases there might be published guidelines about what constitutes a clinically meaningful improvement (see e.g. Sloman, Wruble, Rosen & Rom, 2006), in other cases this might require

the involvement of a clinical expert on the topic. For more information on clinical significance, see Crosby, Kolotkin & Williams (2003) and Lambert & Bailey (2012).

Pre-registration

Pre-registration is the act of depositing the research plan and research hypotheses in a trial registry or other repository before data collection is started. This deposited research plan must be available for other researchers, to help them assess the similarities and differences of the pre-registration and the post-data collection report. (Pre-registration should not be confused with "Registered Reports," where the manuscript is submitted to a journal, peer reviewed, and accepted for publication before data collection starts, or publishing a trial protocol, where the research protocol is published in a journal as a separate paper before data collection stars) (for additional insight and context, see Nosek et al., (2018).

Pre-registration is one of the best practice methodological tools recommended to mitigate researcher- and publication-bias (Nosek et al., 2018). The Hypnosis Efficacy Task Force realizes the usefulness of pre-registration and recommends its use in laboratory and clinical trials. There are two main reasons for the exclusion of pre-registration in the current efficacy guidelines, both of which stem from the fact that pre-registration is a relatively new tool in the fields of medicine and social sciences. Firstly, this means that there is not yet enough data regarding the impact of pre-registration on researcher and publication biases. Secondly, most studies establishing the efficacy of clinical applications of hypnosis were conducted at a time when pre-registration was not yet a standard research practice. Nevertheless, this might soon change since there is a clear trend in the literature in biomedicine and social sciences to treat pre-registration as a standard requirement for confirmatory research and more and more journals include this in their

submission criteria. In future revisions of the recommendations the Hypnosis Efficacy Task Force plans to revisit this issue. Until then, the Hypnosis Efficacy Task Force advocates strongly for the pre-registration of new studies and will regard pre-registration as a marker for reduced risk of bias.

Specificity of the Hypnosis-based Treatment

As stated in the guidelines, the assessment of hypnotizability is not required for establishing efficacy of a hypnosis-based treatment. The Hypnosis Efficacy Task Force notes that it is important to establish that there is a correlation between a treatment effect and hypnotizability. Such a correlation is informative as it can provide valuable information regarding whether the effect is attributable to suggestion or another factor (e.g., motivation). Bowers' doctrine, for example, holds that any effect that is not related to hypnotizability should not be labeled as a hypnotic effect (Woody & Barnier, 2008). However, establishing such a correlation is not a necessary requirement for a treatment to be deemed efficacious. Rather, efficacy is a property of the treatment package as a whole and does not require specificity to any mechanism. The specificity of the treatment is not of primary concern unless alternative treatments have a better benefit to cost ratio. In fact, a meta-analysis by Montgomery and colleagues revealed that the relationship between hypnotizability and treatment outcomes was small (Montgomery et al., 2011). Nevertheless, it is recommended to assess hypnotizability in clinical trials of hypnosis-based treatments to facilitate understanding about the underlying mechanisms.

Blinding

Blinding (masking) of participants/patients and research staff administering the treatment is often considered a key aspect in medical trials to minimize bias due to expectancy and establish that the treatment effect is specific to the proposed effective component of the treatment, for example the specific drug (Shadish et al., 2002). However, as mentioned above, specificity is less of a concern when establishing efficacy. It is true that specificity can affect the costs of a treatment. For example, if the active drug component turns out to be inert and the effect is mainly due to response expectancy, costs can be reduced. However, expectancy plays a central role in psycho-social interventions such as hypnosis (Kirsch, 1994; Kirsch, 2005), and it can be thought of as an active ingredient. Accordingly, the use of classic double-blind placebo-controlled designs from clinical medicine are controversial and difficult to apply in this field (Kirsch, 2005; Parloff, 1986). Nevertheless, certain types of blinding of participants can still be possible using minimally effective control conditions (Jensen & Patterson, 2005) and even sham conditions (Barton et al., 2017; Kendrick et al., 2013; Sliwinski & Elkins, 2013), which might be beneficial in mitigating some experimenter biases and demand-biases, and also in obtaining a better understanding of the role of expectancy in the treatment effect.

On the other hand, blinding of other people involved in the study such as data collectors, outcome assessors, and data analysts is recommended to reduce experimenter biases, and the absence of proper blinding of these individuals should be considered in the risk of bias assessment.

Which Interventions Can Be Considered as Hypnosis-based Treatments?

What types of interventions can, or should a researcher include in a systematic review when conducting efficacy assessment of hypnosis-based treatments? Even though there have

been multiple attempts at defining hypnosis (e.g. Green et al., 2005; Wagstaff, 1998; Elkins et al., 2015), the boundaries are still unclear about what can and cannot be called a hypnosis-based intervention. For example, interventions such as guided imagery, autogenic training, therapeutic suggestions, and Ericksonian conversation etc. may be considered hypnosis-based treatments by some, but not by others. This introduces a certain amount of degrees of freedom for researchers in their inclusion criteria for studies. These degrees of freedom can be sources of bias. For example, certain types of treatments might be included because there are studies with good reported effects in the literature, whereas others might be excluded because of poor results, resulting in an overestimation of the effect size. To overcome this bias, the Task Force has decided to issue a recommendation about what can be regarded as a hypnosis-based treatment for the purposes of efficacy assessment of hypnosis-based interventions. We wanted to allow for as much researcher flexibility as possible whilst remaining responsive to changes in the field regarding the definition of hypnosis and still limiting possibilities for result-based sampling bias. Accordingly, we recommend that for a study to be included in the efficacy assessment review, the intervention used needs to have been labeled in the paper as “hypnosis” or a close synonym (e.g., “hypnotic treatment,” “hypnotherapy”). In addition, the intervention should align with the current consensus among experts about what can be categorized as hypnosis (for a consensus-based definition, see, e.g., Elkins, Barabasz, Council, & Spiegel, 2015). An intervention can be considered a hypnosis-based intervention even if hypnosis is an adjunct to another intervention, as long as at least one part of the complex intervention is identified as hypnosis by the authors of the paper and it meets the above criteria.

Importantly, this recommendation does not specify how an intervention should be presented to the participants of a study. So, even if the intervention is not labeled directly as hypnosis to the participants, but the study meets the two foregoing criteria, it can be included in the efficacy assessment. Nevertheless, it is important to note that the label used when presenting an intervention to participants influences efficacy, and that the label "hypnosis" seems to have a considerable positive effect (see, e.g., Gandhi & Oakley, 2005; Schoenberger, Kirsch, Gearan, Montgomery & Pastyrnak, 1997). Thus, pooling of studies where the intervention is labeled as hypnosis for participants with other studies where other labels are used is discouraged, since it is likely that interventions with the hypnosis label will have larger effect size. Rather, studies with different labels can be treated separately, and this factor can be included in a moderation analysis in the meta-analysis. The recommendations of the Task Force could be useful for researchers who want to make clinical recommendations for other types of interventions similar to hypnosis, such as guided imagery, autogenic training, therapeutic suggestions, and Ericksonian conversation.

Importance of Independent Replication

Independent replication is held as the gold standard for verifying the reliability of scientific claims (Frank & Saxe, 2012). Recent large-scale replication efforts indicate that only about 50% of findings reported in the top tier journals of psychological science are reproducible, even with the direct involvement of the original authors (Baker, 2015; Camerer et al., 2018; Open Science Collaboration, 2015; Owens, 2018). This demonstrates that it is unwise to base practical recommendations on a single research report, however prestigious the journal it was reported in. Thus, the Task Force recommends that the highest level of evidence rating should

only be issued for clinical applications that have been demonstrated to be effective by at least two independent research teams, or by at least one multi-center clinical trial.

Treatment Fidelity

Treatment fidelity means that the intervention is executed consistently as intended by all intervention deliverers (therapists) in the study. This is an important aspect of clinical research that can have a great influence on the effectiveness of the intervention measured in the study. Treatment fidelity can be increased through training and supervision of therapists, clear and comprehensive treatment manuals, and using intervention protocols that are easy to execute consistently. Furthermore, the experience level and allegiance of the therapist to the interventions used can also influence as-intended treatment implementation. The Task Force highly recommends reporting this information in papers on individual clinical trials. Ideally, treatment fidelity should be taken into consideration during the efficacy assessment, and studies with demonstrated treatment fidelity should be weighted higher than other studies or studies where problems are identified in treatment fidelity. However, currently the transparent reporting of these factors is very uncommon in research papers, so it is hard to incorporate these in the efficacy assessment process today. That is why the list of recommendations do not include this aspect currently. The new Cochrane Risk of Bias Tool (RoB 2) incorporates a new risk of bias category “Bias due to deviation from intended interventions” where intervention fidelity is taken into account especially when blinding of participants and therapists is not possible (Munder & Barth, 2018; Sterne, et al. 2019). Since the RoB 2 is a part of the GRADE assessment, reviewers can already incorporate issues related to deviations from intended protocols into their efficacy

assessment. As the reporting of factors contributing to treatment fidelity will become more common, the Task Force may include a recommendation regarding this aspect of clinical trials.

Summary

The Hypnosis Efficacy Task Force was assembled in 2018, with the collaboration of six North American and international hypnosis organizations. Nine leaders in the field of hypnosis participated in monthly Task Force meetings from 2019 through 2021, and five additional leaders in the field reviewed the deliberations and recommendations of the Task Force and provided guidance.

The Hypnosis Efficacy Task Force focused on three objectives: 1) Developing a set of guidelines for the assessment of the efficacy of hypnosis applications, based on methodological criteria, 2) Formulating recommendations for best practices in future outcomes research on clinical hypnosis, and 3) Conducting an international survey of clinicians, researchers, and students in the field of hypnosis, to provide the most comprehensive picture to date on current practices and views.

This report addresses the first objective, creating guidelines for the assessment of efficacy. The Hypnosis Efficacy Task Force recommends that any researcher assessing the efficacy of hypnotic interventions for a specific hypnosis application utilize a well-documented and widely respected evidentiary standard such as the GRADE system. This report suggests several adaptations of the GRADE standards for hypnosis research, based on challenges specific to the study of hypnosis. The Task Force report includes ten specific Guidelines for the Assessment of Efficacy of Clinical Hypnosis Applications.

Finally, the report addresses several recurrent issues in hypnosis research: the use of the GRADE system, the value of pre-registration of research protocols, the value of including an assessment of hypnotizability in outcome research, the challenges of blinding in hypnosis research, the question of which interventions as hypnosis in outcome research, and the importance of independent replication in outcome research.

Two additional papers will be forthcoming from the Task Force, the first reporting the Task Force recommendations for best practices in future outcomes research in hypnosis, and the second summarizing the results of the international survey of hypnosis researchers and practitioners.

Disclosures

Authors contributions

ZK drafted the guidelines based on the Task Force meetings, was involved in coordinating Task Force discussions on the guidelines and led manuscript write-up.

DM initiated the Task Force, was involved in coordinating Task Force discussions on the guidelines and participated in manuscript write-up.

GDB, GE, DBT, KV, OP, PDS and PJW attended the meetings of the Task Force on the guidelines, contributed to formulating and editing the guidelines, reviewed and approved the final wording of the guidelines, and participated in writing the manuscript.

Conflict of interest

Several members of the Task Force are involved in clinical research and have developed hypnosis-based treatments for various clinical applications. The guidelines listed in this paper

may affect how their research findings and the interventions developed by them will be evaluated. The members of the Task Force were aware of these conflicts of interest during their work and made every effort to mitigate these and be objective.

Funding

Zoltan Kekecs was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Science. Katalin Varga was supported by MTA-ELTE Lendület Adaptation Research Group, Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary. Saybrook University, Pasadena, CA, provided the ZOOM platform for Task Force meetings.

Acknowledgement

We want to acknowledge the support of the external consultants of the Task Force: Mark Jensen, Elvira Lang, and David Patterson, who reviewed the list of guidelines and recommended improvements.

Prior versions

A previous version of the manuscript was posted on PsyArXiv as a preprint: https://psyarxiv.com/2cqaw

References

Andrews, J., Guyatt, G., Oxman, A. D., Alderson, P., Dahm, P., Falck-Ytter, Y., Nasser, M., Meerpohl, J., Post, P. N., Kunz, R., Brozek, J., Vist, G., Rinf, D., Akl, E. A., & Schünemann, H. J. (2013). GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *Journal of Clinical Epidemiology, 66*(7), 719-725. doi:10.1016/j.jclinepi.2012.03.013

Atkins, D., Eccles, M., Flottorp, S., Guyatt, G. H., Henry, D., Hill, S., Liberati, A., O'Connell, D., Oxman, A. D., Phillips, B., Schünemann, H., Edejer, T. T.-T., Vist, G. E., Williams, J. W., GRADE Working Group. (2004). Systems for grading the quality of evidence and the strength of recommendations. I. Critical appraisal of existing approaches. The GRADE Working Group. *BMC Health Services Research, 4*(1), 38. doi:10.1186/1472-6963-4-38

Baker, M. (2015, April 15). First results from psychology's largest reproducibility test. *Nature: International Weekly Journal of Science*. doi.10.1038/nature.2015.17433

Balshem, H., Helfand, M., Schunemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., Vist, G. E., Falck-Ytter, Y., Meerpohl, J., Norris, S, Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology, 64*(4), 401-6. doi.0.1016/j.jclinepi.2010.07.015

Barber, T. X. (1976). Pitfalls in human research: Ten pivotal points. Pergamon Press.

Barton, D. L., Schroeder, K. C. F., Banerjee, T., Wolf, S., Keith, T., & Elkins, G. (2017). Efficacy of a biobehavioral intervention for hot flashes: A randomized controlled pilot study. *Menopause, 24*(7), 774. doi.10.1097/GME.0000000000000837

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmeid, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikenstein, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human*

Behaviour, 2(9), 637-644. doi.org/10.1038/s41562-018-0399-z

Carlson, L. E., Toivonen, K., Flynn, M., Deleemans, J., Piedalue, K. A., Tolsdorf, E., & Subnis, U. (2018). The role of hypnosis in cancer care. *Current Oncology Reports, 20*(12), 1-9. doi:10.1007/s11912-018-0739-1

Catsaros, S., & Wendland, J. (2020). Hypnosis-based interventions during pregnancy and childbirth and their impact on women's childbirth experience: A systematic review. *Midwifery, 84,* 102666. doi:10.1016/j.midw.2020.102666

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7-18. doi:10.1037//0022-006x.66.1.7

Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. Journal of clinical epidemiology, 56(5), 395-407. doi: 10.1016/S0895-4356(03)00044-1

Elkins, G. R., Barabasz, A. F., Council, J. R., & Spiegel, D. (2015). Advancing research and practice: The revised APA Division 30 definition of hypnosis. *International Journal of Clinical and Experimental Hypnosis, 63*(1), 1-9. doi:10.1080/00207144.2014.961870

Evidence-Based Medicine Working Group. (1992), Evidence-based medicine: A new approach to teaching the practice of medicine. Journal of the American Medical Association, 268(17), 2420-2425. doi:10.1001/jama.1992.03490170092032

Fisch, S., Brinkhaus, B., & Teut, M. (2017). Hypnosis in patients with perceived stress–a systematic review. *BMC Complementary and Alternative Medicine, 17*(1), 1-12.

doi.10.1186/s12906-017-1806-0

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*(6), 600-604. doi:10.1177/1745691612460686

Gandhi, B., & Oakley, D. A. (2005). Does 'hypnosis' by any other name smell as sweet? The efficacy of 'hypnotic' inductions depends on the label 'hypnosis'. Consciousness and Cognition, *14*(2), 304-315. doi:10.1016/j.concog.2004.12.004

Green, J. P., Barabasz, A. F., Barrett, D., & Montgomery, G. H. (2005). Forging ahead: The 2003 APA Division 30 definition of hypnosis. *International Journal of Clinical and Experimental Hypnosis*, *53*(3), 259-264. doi: 10.1080/00207140590961321

Guyatt, G. H., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., DeBeer, H., Jaeschke R., Rind, D., Meerpohl, J., Dahm, P., & Shünemann, H. J. (2011a). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology, 64(4), 383-394. doi:10.1016/j.jclinepi.2010.04.026

Guyatt, G. H., Oxman, A. D., Kunz, R., Atkins, D., Brozek, J., Vist, G., Alderson, P., Glasziou, P., Falck-Ytter, Y., & Shünemann, H. J. (2011b). GRADE guidelines: 2. Framing the question and deciding on important outcomes. Journal of Clinical Epidemiology, 64(4), 395-400. doi:10.1016/j.jclinepi.2010.09.012

Guyatt, G. H., Oxman, A. D., Kunz, R., Falck-Ytter, Y., Vist, G. E., Liberati, A., & Schünemann, H. J. (2008a). Going from evidence to recommendations. British Medical Journal, 336(7652), 1049-1051. doi:10.1016/j.jclinepi.2012.03.013

Guyatt, G. H., Oxman, A. D., Kunz, R., Vist, G. E., Falck-Ytter, Y., & Schünemann, H. J. (2008b). What is "quality of evidence" and why is it important to clinicians? British Medical Journal, 336(7651), 995-998. doi:10.1136/bmj.39490.551019.BE

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008c). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. British Medical Journal, 336(7650), 924-926. doi:10.1016/j.jclinepi.2010.07.017

Holman, L., Head, M. L., Lanfear, R., & Jennions, M. D. (2015). Evidence of experimental bias in the life sciences: Why we need blind data recording. PLoS Biology, 13(7), e1002190. doi:10.1371/journal.pbio.1002190.

Jensen, M. P., Jamieson, G. A., Lutz, A., Massoni, G., McGeown, W., Santarcangelo, E. L., Demertzi, A., DePascalis, V., Banyai, E. I., Rominger, C., Vuilleumier, P., Faymonville, M.-E., & Terhune, D. B. (2017). New directions in hypnosis research: strategies for advancing the cognitive and clinical neuroscience of hypnosis. Neuroscience of Consciousness, 3(1), nix004. doi:10.1093/nc/nix004

Jensen, M. P., & Patterson, D. R. (2005). Control conditions in hypnotic-analgesia clinical trials: challenges and recommendations. *International Journal of Clinical and Experimental Hypnosis*, *53*(2), 170-197. doi.10.1080/00207140590927536

Kendrick, C., Koep, L., Johnson, A., Fisher, W., & Elkins, G. (2013). Feasibility of a sham hypnosis: Empirical data and implications for randomized trials of hypnosis.

Contemporary Hypnosis and Integrative Therapy, 29(4), 317-331.

Kendrick, C., Sliwinski, J., Yu, Y., Johnson, A., Fisher, W., Kekecs, Z., & Elkins, G. (2016). Hypnosis for acute procedural pain: A critical review. *International Journal of Clinical and Experimental Hypnosis, 64*(1), 75-115. doi:10.1080/00207144.2015.1099405

Kirsch, I. (1994). Clinical hypnosis as a nondeceptive placebo: Empirically derived techniques. *American Journal of Clinical Hypnosis, 37*(2), 95-106. doi:10.1080/00029157.1994.10403122

Kirsch, I. (2005). Placebo psychotherapy: Synonym or oxymoron? *Journal of Clinical Psychology, 61*(7), 791-803. doi.0.1002/jclp.20126

Lambert, M. J., & Bailey, R. J. (2012). Measures of clinically significant change. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), APA handbook of research methods in psychology, Vol. 3. Data analysis and research publication (pp. 147–160). American Psychological Association. doi: 10.1037/13621-007

LaVaque, T. J., Hammond, D. C., Trudeau, D., Monastra, V., Perry, J., Lehrer, P., Matheson, D., & Sherman, R. (2002). Template for developing guidelines for the evaluation of the clinical efficacy of psychophysiological evaluations. *Applied Psychophysiology and Biofeedback, 27*(4), 273–281.

Lynn, S. J., Kirsch, I., & Hallquist, M. (2008). Social cognitive theories of hypnosis. In M. R. Nash & A. Barnier (Eds.), The Oxford handbook of hypnosis: Theory, research and practice (pp. 111-140). Oxford University Press.

Madden, K., Middleton, P., Cyna, A. M., Matthewson, M., & Jones, L. (2016). Hypnosis for pain management during labour and childbirth. *Cochrane Database of Systematic Reviews*, 2016(5), CD009356. doi:10.1002/14651858.CD009356.pub3

Montgomery, G. H., Schnur, J. B., & David, D. (2011). The impact of hypnotic suggestibility in clinical care settings. *International Journal of Clinical and Experimental Hypnosis*, *59*(3), 294-309. doi.10.1080/00207144.2011.570656

Munder, T., & Barth, J. (2018). Cochrane's risk of bias tool in the context of psychotherapy outcome research. Psychotherapy Research, 28(3), 347-355. doi:10.1080/10503307.2017.1411628

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The pre-registration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600-2606. doi:10.1073/pnas.1708274114

OCEBM Working Group. (2011). The Oxford 2011 Levels of Evidence Table. Oxford Centre for Evidence-Based Medicine. https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251): 11c4716. doi:10.1126/science.aac4716

Owens, B. (2018, November 19). Replication failures in psychology not due to differences in study populations. *Nature*. doi.org/10.1038/d41586-018-07474-y

Parloff, M. B. (1986). Placebo controls in psychotherapy research: A sine qua non or a placebo

for research problems? *Journal of Consulting and Clinical Psychology, 54*(1), 79-87. doi:10.1037//0022-006x.54.1.79

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence-based medicine: What it is and what it isn't. British Medical Journal, 312 (7023), 71-72. doi:10.1136/bmj.312.7023.71

Schoenberger, N. E., Kirsch, I., Gearan, P., Montgomery, G., & Pastyrnak, S. L. (1997). Hypnotic enhancement of a cognitive behavioral treatment for public speaking anxiety. *Behavior Therapy, 28*(1), 127-140. doi: 10.1016/S0005-7894(97)80038-X

Schünemann, H., Brozek, J., Guyatt, G., & Oxman, A. (Eds.) (2013). GRADE Handbook – Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach (updated October 2013). GRADE Working Group. Available from https://gdt.gradepro.org/app/handbook/handbook.html

Schünemann, H., Vist, G., Higgins, J., Santesso, N., Deeks, J., Glasziou, P., Akl, E., & Guyatt, G. (2021). Interpreting results and drawing conclusions. In J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), Cochrane handbook for systematic reviews of interventions version 6.2 (updated February 2021). Retrieved July 8, 2021 from: https://training.cochrane.org/handbook/current/chapter-15

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference. Chapter 3: Construct validity and external validity. (pp. 64-102).* Houghton Mifflin.Siemieniuk, R., & Guyatt, G. (2021). What is GRADE? BMJ Best Practice. https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-

is-grade/.

Sliwinski, J., & Elkins, G. R. (2013). Enhancing placebo effects: Insights from social psychology. *American Journal of Clinical Hypnosis, 55*(3), 236-248. doi:10.1080/00029157.2012.740434.

Sloman, R., Wruble, A. W., Rosen, G., & Rom, M. (2006). Determination of clinically meaningful levels of pain reduction in patients experiencing acute postoperative pain. Pain Management Nursing, 7(4), 153-158. doi: 10.1016/j.pmn.2006.09.001.

Sterne, J. A., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., ... & Higgins, J. P. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ, 366:l4898. doi:10.1136/bmj.l4898.

Tan, G., Shaffer, F. Lyle, R., & Teo, I. (Eds.), (2016). *Evidence-based treatment in biofeedback and neurofeedback* (3rd ed.). Association for Applied Psychophysiology and Biofeedback.

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically supported treatment: Recommendations for a new model. Clinical Psychology: Science and Practice, 22(4), 317–338. doi: 10.1037/h0101729.

Wagstaff, G. F. (1998). The semantics and physiology of hypnosis as an altered state: Towards a definition of hypnosis. *Contemporary hypnosis*, *15*(3), 149-165.

Woody, E. Z., & Barnier, A. J. (2008). Hypnosis scales for the twenty-first century: What do we know and how should we use them? In M. Nash & A. J. Barnier (Eds.), *The Oxford*

handbook of hypnosis: Theory, research and practice (pp. 255-281). Oxford University Press.

Table 1.

Quality of evidence	Meaning
Very low	The true effect is probably markedly different from the estimated effect
Low	The true effect might be markedly different from the estimated effect
Moderate	The authors believe that the true effect is probably close to the estimated effect
High	The authors have a lot of confidence that the true effect is similar to the estimated effect

Table 2.

Strength of recommendation	Factors considered during the recommendation
Strong	- The desirable effects greatly outweigh the undesirable effects - The quality of evidence is relatively high - The values and preferences of patients related to the desirable and undesirable effects are clear - The cost of treatment is acceptable compared to the risks and benefits involved
Weak	- The difference between the desirable and undesirable effects is not large enough to warrant a strong recommendation - The quality of evidence supporting clinically meaningful beneficial effects is not high enough to warrant a strong recommendation - There is uncertainty about, or variability in, values and preferences related to the weight of desirable and undesirable effects - The cost of treatment is too high compared to the risks and benefits involved

Box 1. GRADE Quality of evidence rating

Box 1. GRADE Quality of evidence rating

There are four levels for the quality of evidence rating in GRADE: very low, low, moderate, and high.

Randomised trials begin as high quality evidence and observational studies as low quality evidence, and this initial rating is upgraded or downgraded based on the factors below:

Quality of evidence may be downgraded due to the following factors:

- Risk of bias: There is evidence for risk of bias in the design of the studies included in the review or other important study limitations.
- Inconsistency: There is considerable heterogeneity in the effects reported by the studies in the review.
- Indirectness: The studies in the review don't include the relevant interventions (only similar interventions), and/or if the studies don't include the populations or outcomes of primary interest.
- Imprecision: There is uncertainty about the size of the effect, for example because the studies include relatively few participants and/or events and thus have a wide confidence interval around the estimate of the effect.
- Publication bias: There is evidence for selective publication of studies resulting in a systematic bias in the effect estimates.

Quality of evidence may be upgraded due to the following factors:

- Large magnitude of an effect: There is reliable evidence that the effects are large (risk ratio over 2).
- Dose-response gradient: There is reliable evidence that an increase in the dose of the intervention leads to an increase in the effects.
- Effect of plausible residual confounding: All residual confounders are expected to decrease the magnitude of the effect. Unmeasured determinants or moderators of the effect (residual confounders) can be distributed unequally between intervention and control groups in observational studies. In some cases an observational study is conducted in a way that only such unaccounted confounders remain that would result in an underestimate of an apparent treatment effect.

For more details on how the above mentioned factors should be used in determining quality of evidence, see the Quality of evidence section in the GRADE Handbook. (Schünemann, Brozek, Guyatt & Oxman, 2013).