The Future of Professional Hypnosis:
Comment on Kirsch, Mazzoni, and Montgomery

Edward J. Frischholz
Rush North Shore Medical Center

Abstract

I believe the paper by Kirsch, Mazzoni and Montgomery (this issue) should surprise about 95% of ASCH members (maybe only 93% of SCEH members) because the three facts espoused in their paper speciously seem to be 100% true. To paraphrase from their abstract: 1) nothing that can be produced by hypnotic induction plus suggestion cannot also be produced by suggestion alone; 2) administration of a hypnotic induction does not produce a meaningful increase in response to suggestion relative to suggestion alone; and 3) responsivity to suggestions are highly correlated to responsivity on the same measure when preceded by a hypnotic induction ceremony. In order to persuade that these propositions are true, several objections to them must be addressed. However, just because one’s facts are true does not mean that one’s interpretation of the facts and their interrelationships are also true. The ramifications of the above facts and their interrelationships for the future of professional hypnosis (experimental, clinical and forensic) are identified and discussed.

Keywords: Hypnosis, hypnotizability, suggestion, suggestibility.
Commentary

The Future of Hypnosis:
Comment on Kirsch, Mazzoni and Montgomery

Kirsch, Mazzoni and Montgomery (this issue) seem to allege that the field of professional hypnosis has developed an amnesia for some basic facts which have emerged over the last 75 years using a research design originally developed by Clark Hull (Hull, 1933) but often speciously attributed to Hilgard (Hilgard, 1965). They propose to reverse this amnesia by reminding the field of three basic findings: 1) nothing that can be produced by a hypnotic induction ceremony plus suggestion cannot also be produced by suggestion alone; 2) administration of a hypnotic induction ceremony does not produce a meaningful increase in suggestibility relative to baseline suggestibility response; and 3) baseline suggestibility scores are highly correlated with scores on the same scale taken after the administration of a hypnotic induction ceremony. All three of these alleged facts seem to be myths that never were true, but always will be in the mind of some theorists. Now that we are all awake and aware that the same facts may have vastly different interpretations, we can seriously consider their ramifications for the future of professional hypnosis.

Suggestion Alone vs. Hypnotic Induction Plus Suggestion

Two Californians were contemplating a trip to Florida while standing under a full moon. One of them queried, “Which do you think is farther away, Florida or the moon?” The other one responded, “Well Duh! You can see the moon!” This example aptly illustrates that appearances can be deceiving and misdirections can appear appealing. It has direct relevance to understanding the nature of hypnosis, hypnotic responsivity, suggestion, and suggestibility.

The Moon will always appear to be closer than a location 100 miles away, but it will still always be farther away. Likewise, both the field of professional hypnosis and the lay public usually seem impressed with the dramatic “domain” of hypnotic phenomena, e.g., involuntary motor movements, motor paralyses, amnesia, positive and negative hallucinations, age regression, etc. (Hilgard, 1973). But these are not universal responses to either being administered a hypnotic induction ceremony or baseline suggestibility. Data collected in Hilgard’s laboratory indicates that approximately 45% of subjects earned a score of 4 or less on the Stanford Form C scale and were categorized as “low” in terms of their general level of hypnotic responsivity (Hilgard, 1965, p. 236). In contrast, about 5% of subjects earned scores greater than 10 on the 12 point Form C scale and were categorized as “very high” (i.e., likely to manifest most of the “domain” of hypnosis).

While moon-gazing at the “domain” of hypnotic phenomena, we have lost sight of the fact that they are not the mean response observed in the average subject. To be sure, they are interesting, but I believe that it is just as important to explain why such dramatic responses are not produced by suggestion (hypnotic or waking) in the majority of subject’s studied. In other words, I believe any theory of hypnosis must explain the wide variation of individual differences in response to suggestion following a hypnotic induction ceremony or baseline suggestibility. Why do a small, but significant, percentage of subjects respond to suggestion (with or without a hypnotic induction) while the majority of subjects do not?

Hypnotic Induction Ceremonies Produce No Meaningful Gains in Responsivity to Suggestion

The noted Chicago social scientist Al Capone once remarked, “you can get a lot farther with kindness and a gun than you can with kindness alone.” Capone’s common
sense wisdom also had great metaphoric relevance for understanding the nature of hypnosis and suggestion. I bet that a Capone hypnotic induction (hypnotic induction plus machine guns) would produce a highly significant increase in responsivity to baseline suggestion response relative to just a hypnotic induction with no machine guns. But, all kidding aside, no Institutional Review Board is ever likely to approve a study with such a design.

Instead, I examine data from the classic Weitzenhoffer & Sjoberg (1961) experiment on waking-hypnotic suggestibility, particularly the scatterplot of their data points which was reproduced in Hilgard’s book (1965, p. 29). (Please note that one subject’s data point seems to be missing and there are only 43 subjects, instead of the alleged 44, accounted for in the scatterplot). This classic study assessed the purported gain achieved (i.e., generalized hyper-suggestion response) when one examines difference scores in suggestion response (measured on the same exact scale) between waking suggestibility and hypnotic induction conditions. An overall significant gain in response of 3.15 items ($sd = 2.35$) was observed using a within subjects design analysis (waking vs. hypnotic induction).

However, if one analyzes the data as a mixed, between subjects (initial suggestion response two items or less passed vs. greater than three or more items passed) and within subjects (waking vs. hypnotic induction) design, a more complex interpretation of the data is required (Sheehan & Perry, 1976). Although, the within subjects factor is still significant, it is qualified by the also significant between subjects factor of initial suggestion response scores. Thirty one (or 72%) of the sample earned a baseline suggestibility score of two or less and their gain in suggestion responsivity was an additional 2.00 items passed ($sd = 2.49$). In contrast, of the remaining twelve subjects (28%) with a waking suggestibility score of greater than two, an average gain of 3.25 items passed ($sd = 2.42$) occurred. In other words, a significant increase in response to baseline suggestibility was observed, but the increase was also significantly greater in subjects with high baseline suggestion response compared to subjects with low or no baseline suggestion response.

Two considerations should be taken into account when interpreting the kind of data from the Weitzenhoffer & Sjoberg (1961) study and newer studies which addressed the same issue. First, one must distinguish between a statistically significant finding and the size of the effect observed (Cohen, 1977). Certainly, the gains observed in the Weitzenhoffer & Sjoberg (1961) study are statistically significant and have an effect size increase of more than a full standard deviation. However, other studies have not always replicated the size or the significance of the hypnotic induction effect. For example, Katz (1979) found that two of his “social learning” hypnotic inductions produced significant increases in hypnotic responsivity relative to baseline. But, his third induction, a ‘traditional sleep-trance’ induction produced no significant increase relative to baseline, yet this induction was similar to that used in the Weitzenhoffer & Sjoberg (1961) study. Other studies (Braffman & Kirsch, 1999) have also failed to find significant increases or much smaller increases than a full standard deviation in size.

Second, the hypothesized hypnotic vs. waking increase was predicted to be much larger than those actually observed and was categorized as a generalized increase or state of hyper-suggestibility (Bernheim, 1888; Hull, 1933). If this were so, one would expect tremendous mean increases on the hypnotic posttest relative to the waking pretest and a substantial reduction in variation on the posttest. The cumulative data rarely, if ever, produce extremely large increases in responsivity and almost never yield significantly reduced variation in posttest scores. Thus, one would concur with Kirsch et al (this issue) that hypnotic induction ceremonies rarely produce meaningful increase in responsivity relative to baseline levels of suggestibility.
Waking Suggestibility is Highly Correlated with Hypnotic Suggestibility (especially when the same scale is used)

There is an old adage which goes something like this: “If it walks like a duck, has feathers like a duck, smells like a duck and even quacks like a duck…..then it’s a duck.” Likewise, if subjects respond to suggestion in a waking (e.g., baseline, imagination) context, then studies have shown that suggestibility scores following the administration of a hypnotic induction ceremony will be highly correlated with waking suggestibility scores (especially if the same scale is used (Barber, 1969; Frischholz, Blumstein & Spiegel, 1982; Hilgard, 1965; Weitzenhoffer & Sjoberg, 1961), but there are many qualifications to the above observation. For example, Katz (1979) hypothesized that constructing more powerful hypnotic inductions would lead to greater hypnotic responsivity, which in turn should hypothetically increase treatment responsibility. He randomly assigned subjects who had been prescreened on a Stanford hypnosis scale (under an imagination baseline condition) to three groups. Two of the groups received either a “social learning” hypnotic induction or a “social learning” hypnotic induction plus progressive muscle relaxation. The third group received a ‘traditional sleep-trance’ induction. All were then readministered the same Stanford hypnosis scale. Using a one way analysis of variance design and post hoc analyses, both social learning groups showed a significant increase from baseline of about three additional items passed similar to the Weitzenhoffer and Sjoberg (1961) study. In contrast, the ‘traditional sleep-trance’ group evidenced no significant increase from baseline.

Frischholz, Blumstein & Spiegel (1982) reanalyzed the Katz (1979) data using a one way analysis of covariance design with baseline imagination suggestion response as the covariate and posttest scores as the dependent variable. Results still indicated a significant hypnotic induction effect (accounting for about 17% of the variation in posttest scores). However, this significant treatment effect was overshadowed by the much more powerful covariate effect which accounted for approximately 50% of the variance in posttest scores. No covariate by treatment interaction was observed.

Frischholz et al (1982) concluded that this “illustrates what is often overlooked when reporting results on experiments that attempt to enhance hypnotic responsivity relative to a pretreatment baseline; that is, pretest scores tend to be significantly correlated with post scores (Perry, 1977). This is because a particular individual’s absolute standing on the dependent measure may vary as a result of being exposed to a particular treatment. However, this person’s relative within-group standing tends to remain similar to his pretreatment standing on the same dependent measure.” (Frischholz, Blumstein & Spiegel, 1982, p. 768).

Now that I have identified the statistical complexity of the issue, let’s get our ducks in a row about this matter and other considerations. First, from a purely statistical perspective, high test-retest correlations can obtain even when there are substantial mean differences between the two test administrations. For example, suppose an eight item baseline suggestibility measure (like the Barber Suggestibility Scale) was administered to a group. Wide individual differences in responsivity would likely be found (Barber, 1969). Next, let’s say a hypothetically powerful hypnotic induction was administered to this group and the scale re-administered. If one were to add a constant score increase to each individual of 5 score points, then the baseline score would be perfectly correlated with the posttest score, despite the highly significant mean increase of 5 score points. If one only looked at the perfect correlation, one might miss the dramatic mean differences between the pre and posttests.

Second, when considering whether increased baseline suggestibility is due to the
administration of a hypnotic induction ceremony, one should first look at the least common denominator in this equation—suggestion and baseline suggestibility. There is no consensus about a theoretical or operational definition of suggestion or suggestibility (Frischholz, 2002; Frischholz, 2005; Gudjonsson, 2003; Hull, 1933; Stukat, 1958). Many different theories of suggestion and suggestibility have been offered and response to different suggestibility scales, although sometimes significantly intercorrelated, still indicate that each possess unique variance when compared to other suggestibility scales. For example, a recent study by Polczyk and Pasek (2006) found no significant correlation between the Group Sensory Suggestibility Scale (an alleged measure of indirect suggestibility) and the Gudjonsson compliance scale (Gudjonsson, 2003). However, when used as predictors of the Barber Suggestibility Scale (Barber, 1969), the combination of both the Group Sensory Suggestibility Scale and the Gudjonsson Compliance Scale as two independent predictors in a multiple regression paradigm, resulted in a higher (multiple) correlation with the Barber Suggestibility Scale than either scale did alone. The obvious conclusion is that there are different types of suggestibility and how much they correlate with other measures of suggestibility depends on whether they are used alone or in a multiple regression analysis. If suggestibility is used as the operational indicator of suggestion, then there are also probably different types of suggestion as well.

Likewise, Gudjonsson has developed a suggestibility scale (Gudjonsson Suggestibility Scale (GSS); Gudjonsson, 2003), which he proposes measures a construct called interrogatory suggestibility (or the tendency to respond to misleading questions about an event by later incorporating this misleading information into one’s recall of that event at a later time). Six studies (Gordon, Gwynn, & Spanos, 1993; Gwynn & Spanos, 1996; Gwynn, Spanos, Nanoco, & Chow, 1995; Hardarson, 1985; Register & Kihlstrom, 1988; Young, Bentall, Slade, & Dewey, 1987) using a variant of the GSS and such criterion measures as the Barber Suggestibility Scale, the Carleton University Responsiveness to Suggestion Scale or the Harvard Group Scale found no significant correlation between the various measures of interrogatory suggestibility and the various measures of hypnotic responsivity. In contrast, two other studies (Linton & Sheehan, 1994; Sheehan, Garnett, & Robertson, 1993) did find a significant (but low effect size) relationship between some GSS scores and the Harvard Group Scale of Hypnotic Susceptibility. Nevertheless, the cumulative nature of the data indicates that interrogatory suggestibility as measured by GSS scores is not highly correlated with hypnotic responsivity as it is measured by the Barber Suggestibility Scale, the Carleton University Responsiveness to Suggestion Scale or the Harvard Group Scale. Collectively, the data shows weak, but sometimes significant, relationships between suggestibility as measured by different scales when an attempt is made to correlate it with other scales which usually have been taken to measure hypnotic responsivity (or at least suggestibility preceded by the administration of a hypnotic induction).

Finally, one must consider the alleged baseline condition from which we measure increases (or decreases) in the posttest measure. Is simply administering the pretest with no other instructions the same thing as administering it under imagination instructions? I think not.

**Ramifications and Conclusions**

It should be abundantly clear since the last two years that there is strong disagreement about both theoretical and operational definitions of hypnosis (Araoz, 2005; Barabasz, 2005; Daniel, 2005; Green, Barabasz, Barret, & Montgomery, 2005; Hammond, 2005; Heap, 2005; McConkey, 2005; Nash, 2005; Rossi, 2005; Spiegel & Greenleaf, 2005,
Woody & Sadler, 2005; Yapko, 2005). This situation is no different than the time of Hull, Hilgard, Bernheim or perhaps even back to Mesmer (Barber & Glass, 1962; Frischholz, 2005; Hull, 1933, Hilgard, 1965; Spiegel & Spiegel, 1978) perhaps back to Mesmer. While we seem to have come to some agreement about what to call individual differences in “hypnotizability” or “hypnotic responsivity” (Christensen, 2005) since Weitzenhoffer’s (1980) demurral 26 years ago, the field of professional hypnosis cannot agree on what hypnosis is or how to explain individual differences in responsivity to hypnotic inductions.

While unchallenged agreement exists that something called hypnosis exists and have defined an operational “domain” (Hilgard, 1973) of phenomena to study, we often seem to forget that a small percentage of individuals manifest this entire domain behaviorally or subjectively. The majority of subjects evidence, at best, just a few of these phenomena and many do so without being administered a hypnotic induction ceremony. And, the administration of a hypnotic induction ceremony seems to produce no meaningful, dazzling increase (i.e., a state of “hyper-suggestibility better than an additional 2-3 items passed) in responsivity relative to baseline response on the same scale.

Calling responsivity to various scales “suggestibility” seems to have stimulated no theoretical or operational scientific advances in the study of professional hypnosis. Furthermore, while there is sufficient empirical evidence to demonstrate that qualitatively different types of suggestion (e.g., direct vs indirect) and different types of suggestibility (primary, secondary, interrogatory) exist, no serious consideration has been given to the notion that there are qualitatively different types of hypnotic responsivity. While many agree that a state of hypnosis is a conscious state qualitatively different from some type of waking-baseline state, i.e., administration of a hypnotic induction produces a qualitatively different, special conscious state, (Christensen, 2005), little attention has been paid to the proposition that the same behavioral response may be accompanied by a qualitatively different subjective experience. Consider the extreme example of an Al Capone induction. While such an induction is likely to produce most of the “domain” of hypnotic phenomena, it is doubtful that anyone would seriously propose that this behavioral responsivity is accompanied by the same subjective experience in a highly hypnotizable subject administered a hypnotic induction.

Is there any hope for the field of professional hypnosis? I believe so. First, if we cannot agree on what hypnosis is, then perhaps we can agree on what it is not. For example, hypnosis is not a specific type of therapy or memory retrieval technique (Frischholz, 1997; Frischholz, 2000; Frischholz & Spiegel, 1983; Schefflin & Frischholz, 1999). Rather, hypnosis is an adjunct phenomenon that can be utilized with a variety of different primary treatment strategies (e.g., direct suggestion, psychodynamic treatments, behavioral therapy, cognitive-behavioral therapy, EMDR, etc) as well as a variety of different types of memory retrieval strategies. I believe Kirsch, Mazzoni, & Montgomery (this issue) would agree with this conclusion since they have referred to the use of a hypnosis as a therapeutic adjunct in other publications (Kirsch, Montgomery, & Saperstein, 1995; Montgomery, David, Winkel, Silverstein, & Bovbjerg, 2002).

Second, I hope we can agree that the administration of a hypnotic induction does not, by itself, produce dissociative psychopathology (Brown, Frischholz, & Schefflin, 2000). Rather, patients suffering from different types of psychiatric disorders show both quantitative and qualitative differences on various indicators of hypnotic responsivity (Frischholz, Lipman, Braun, & Sachs, 1992).

Third, if Kirsch et al (this issue) are right, then there is nothing special about
hypnosis compared to other social influence phenomena. If that is true, then there is no need for a per se exclusion rule for a person’s testimony taken after the administration of a hypnotic induction. Hypnotic inductions do not promote false memories about an event (AMA, 1985), misleading questions do (Scoboria, Mazzoni, & Kirsch, 2005). Even if there was a thing as a hypnotic “state,” it would have to be empirically shown that such a state was associated with a greater susceptibility to misleading information.

Fourth, there are numerous studies which indicate psychophysiological and neurophysiological differences between those that score high and low on various measures of hypnotic responsivity (Barabasz, Barabasz, Jensen, Calvin, Trevisan, & Warner, 1999; Freeman, Barabasz, Barabasz, & Warner, 2000; Jensen, Barabasz, Barabasz, & Warner, 2001; Kosslyn, Thompson, Costiantini-Ferrando, Alpert, & Spiegel, 2001; Smith, Barabasz & Barabasz, 1996; Spiegel, 1972; Spiegel & Spiegel, 1978/2004). But if there is no agreement about a theoretical or operational definition of hypnosis, there can be no agreement about what is or is not a hypnotic induction ceremony or what is or is not a neurophysiological indicator of its presence.

Fifth, and probably most important, there is broad agreement that there is wide variation in response to suggestion and hypnosis. If we are ever able to explain why such variation exists, we must encourage the systematic measurement of these variations. For example, hypnotizability scores have been empirically shown to predict positive treatment response to cognitive restructuring strategies (Spiegel, Frischholz, Maruffi, & Spiegel, 1981; Spiegel, Frischholz, Fleiss, & Spiegel, 1993). Therefore, I propose that assessing hypnotic responsivity become part of the standard of practice in clinical and forensic hypnosis just as it has in experimental hypnosis.

In order to reverse the suggested posthypnotic amnesia item on the Stanford Form C scale subjects are told, “Now you can remember everything.” I hope that the paper by Kirsch, Mazzoni and Montgomery (this issue) will serve as such a reversal cue for the field of professional hypnosis. There are many facts and interrelationships between such facts which have been established through scientific research over the last century. The field of professional hypnosis needs to remember these facts and come to a consensus about the best way to explain these facts and their interrelationships.

References


Commentary


Commentary


