

Mapping Evidence-Based Treatments for Children and Adolescents: Application of the Distillation and Matching Model to 615 Treatments From 322 Randomized Trials

Bruce F. Chorpita
University of California, Los Angeles

Eric L. Daleiden
Kismetrics, LLC

This study applied the distillation and matching model to 322 randomized clinical trials for child mental health treatments. The model involved initial data reduction of 615 treatment protocol descriptions by means of a set of codes describing discrete clinical strategies, referred to as practice elements. Practice elements were then summarized in profiles, which were empirically matched to client factors (i.e., observed problem, age, gender, and ethnicity). Results of a profile similarity analysis demonstrated a branching of the literature into multiple problem areas, within which some age and ethnicity special cases emerged as higher order splits. This is the 1st study to aggregate evidence-based treatment protocols empirically according to their constituent treatment procedures, and the results point both to the overall organization of therapy procedures according to matching factors and to gaps in the current child and adolescent treatment literature.

Keywords: evidence-based, practice elements, common elements, distillation, matching

The call for an increased focus on issues related to implementation, dissemination, and real-world relevance has grown steadily over the past several years (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; Kazdin, 2008; National Advisory Mental Health Council Workgroup on Child and Adolescent Mental Health Intervention Development and Deployment, 2001; U.S. Department of Health and Human Services [HHS], 1999), and the improved understanding of the defining features, relevance, and fit of psychological treatments for various real-world contexts is now a major goal of national significance. As findings relevant to these aims accumulate, it is also possible to gain insights from the existing evidence base of child mental health outcomes by meta-analytic reviews as well as by methods designed to extend or complement those approaches.

Examining Components of Treatments

For example, it seems reasonable to address the basic question, What features characterize successful (i.e., evidence-based) treatments? This question itself has been asked in many ways by looking at the intensity, focus, or theoretical approach of different treatments. However, there has been as yet no systematic review of the actual clinical strategies that characterize evidence-based treatments across the literature. Addressing this question raises a more general issue of the appropriate level of analysis by which to

examine treatments. Despite a long tradition of reviewing and summarizing treatment outcome research in psychology, few recommendations exist regarding a definitive level of analysis. In early reviews, for example, treatments were often examined as a whole (e.g., Is treatment effective?; Smith, Glass, & Miller, 1980). Early reviews were followed by understandable controversy, in which critics pointed out that considering multiple forms of psychotherapy the same raised numerous problems (e.g., see Beutler, 2002; Luborsky et al., 2002; Rounsaville & Carroll, 2002). More recent traditions have involved the examination of the effects of interventions organized at the level of their theoretical background (e.g., Is cognitive behavioral therapy effective?) or at the level of their specific treatment protocol (e.g., Is the Incredible Years program effective?).

Methodologies of this nature have dominated the field since the emergence of historic monographs by the American Psychological Association and its clinical division nearly 15 years ago (Task Force on Promotion and Dissemination of Psychological Procedures, 1995; Task Force on Psychological Intervention Guidelines, 1995) and continue in the area of child psychology as well (e.g., Silverman & Hinshaw, 2008). Reviews in this tradition offered distinct advantages over previous approaches, in that the rules for knowledge accumulation were better articulated, particularly in terms of defining treatment efficacy. However, a possibly unintended consequence was that the methodology came to emphasize treatment manuals as the optimal unit of analysis (Chambless et al., 1996); in other words, to emphasize that a specific protocol (e.g., Adolescent Coping With Depression; Clarke, Rohde, Lewinsohn, Hops, & Seeley, 1999) is evidence-based, as opposed to either its theoretical family (i.e., cognitive-behavioral therapy for depression) or its assembly of component strategies (e.g., cognitive restructuring, pleasant event scheduling).

One complement to this approach is to define a level of analysis that can be aggregated across protocols and studies, such that

Bruce F. Chorpita, Department of Psychology, University of California, Los Angeles; Eric L. Daleiden, Kismetrics, LLC, Satellite Beach, FL.

Bruce F. Chorpita is president of PracticeWise, LLC, a private behavioral health consulting corporation. Eric L. Daleiden, who provided the coded data for this article, works as a consultant to PracticeWise, LLC.

Correspondence concerning this article should be addressed to Bruce F. Chorpita, Department of Psychology, University of California, Los Angeles, Box 951563, Los Angeles, CA 90095. E-mail: chorpita@ucla.edu

inferences about treatment content can be evaluated empirically. More generally, knowing what specific strategies are common among successful treatments affords an enriched understanding of those treatments and may produce important insights or hypotheses about mechanisms as well as the defining boundaries of intervention types. Accordingly, documenting the component strategies that characterize evidence-based treatments for children is the first major aim of this study.

Association With Treatment Context

It also makes sense, aside from documenting these components, to ask whether particular component practices are associated with specific features of the clinical context (including child characteristics). There is an array of rich findings in this tradition that already highlight the importance of examining treatment by client interactions. For example, the observed advantage of behavior therapy over other treatments for children might not have been detected in a sample collapsing across the lifespan (Weisz, Weiss, Han, Granger, & Morton, 1995). It follows logically that there may be findings that remain undetected by current approaches because of methodologies that (a) collapse levels of a factor together on the basis of superficial or rationally determined similarities (e.g., dialectical behavior therapy and systematic desensitization, two rather different types of behavior therapy, potentially being lumped together as behavior therapy) or (b) examine only main effects (e.g., What works for depression?) or two-way interactions (e.g., What works for depressed adolescents?). We know of no reviews that have systematically examined higher order (three- or four-way) interactions across the treatment outcome literature with respect to identifying what has worked under various conditions and their combinations. Thus, the second major aim of the study was to determine whether the component practices that characterize evidence-based treatments in general are organized differently depending on various context features (e.g., Are evidence-based treatments characterized by unique practices in the context of treating depressed girls of Asian American ethnicity?).

The Distillation and Matching Model

The distillation and matching model (DMM; Chorpita, Daleiden, & Weisz, 2005) was designed to provide a detailed description of strategies characterizing evidence-based treatments and to circumvent some of the problems associated with using manuals as the level of analysis (e.g., *empirical redundancy*, or the inability to aggregate similar findings across the literature) as well as those associated with rationally defined treatment approaches (e.g., investigator-driven inferences about the boundaries of a treatment approach, a problem type, or some other context variable). The methodology uses frequency patterns in practice techniques to guide the empirical construction of a distillation tree that organizes the selected literature according to any number of a priori selected variables of interest (e.g., disorder type, age, ethnicity, etc.). The model is broadly designed to (a) empirically accumulate a map of the treatment practices with favorable treatment outcome data, (b) promote understanding of the underlying data relations between treatment practices and client or context variables, and (c) facilitate hypothesis generation regarding potential prescriptive heuristics to apply to novel situations. Chorpita et al. (2005) demonstrated the

feasibility of the model as applied to 43 randomized trials by using a limited code set of 26 practice element codes.

For the current investigation, we used a larger set of codes with established reliability as applied to a comprehensive survey of 322 randomized trials. In our analyses, we made explicit decisions about (a) defining the boundaries of the literature reviewed (i.e., randomized trials for major mental health disorders for children and adolescents; see below), (b) defining the dimensions along which treatment protocols would be coded (i.e., discrete therapeutic procedures, as opposed to alliance, intensity, duration, engagement, or other treatment dimensions), and (c) selecting factors with the potential to serve as moderators (i.e., problem type, age, gender, and ethnicity). The following review is a documentation of the structure of the child and adolescent treatment outcome literature on the basis of those particular methodological assumptions.

Method

Review Procedures

We sought to identify randomized clinical trials of nonpharmacological treatments targeting anxiety disorders, attention deficit and hyperactivity, autistic spectrum, depression, disruptive behavior, substance use, and traumatic stress. For a study to be included, the majority of participants had to be under 19 years old. We did not evaluate psychosocial interventions for health-related conditions (e.g., childhood obesity, diabetes management) or some of the less common mental health conditions among children and adolescents (e.g., bipolar disorder, eating disorders, tic disorders).

Studies contributing to this review were identified through a combination of strategies, including (a) computerized searches of electronic databases for relevant publications; (b) evaluation of studies reviewed by the American Psychological Association's Task Force on Empirically Supported Psychosocial Interventions for Children, the American Academy of Child and Adolescent Psychiatry Practice Parameters, and other major published scientific literature reviews; (c) personal communication with national scholars; and (d) additional ad hoc nominations from members of the coding team and other professionals. A total of 322 randomized clinical trials that spanned a period of 41 years of research were identified. These studies tested 615 different treatment protocols, which were subjected to coding. Additional description of this review, as well as the accompanying reference list, is available in Chorpita and Daleiden (2007).

Initial coding. All studies were coded using the PracticeWise Clinical Coding System (PracticeWise, 2005), which summarizes multiple variables pertaining to studies, study groups, and treatment protocols. A *study* was defined as a clinical research project in which participants were randomized to different study groups. A *study group* was defined as a set of participants, such as a treatment group or a control group, that was randomized within a study to receive a defined protocol. *Protocols* was defined as the description of the set of treatment operations in which members of a particular study group participated. A single publication could contain multiple studies (e.g., Study 1, Study 2), and a single study was sometimes summarized across multiple publications. Also, in the event that a study explicitly examined categorical moderator effects, and different outcomes were found across the different groups, these groups were divided (if the investigator had not

already done so) to create separate studies for coding. For example, a study that found that Treatment A worked better than Treatment B for boys but not for girls (Group \times Time \times Gender interaction) was coded separately as a boys study, in which the Group \times Time interaction indicated a positive outcome, and a girls study, in which the nonsignificant Group \times Time interaction indicated a null outcome. This occurred in six of the publications coded.

Each study and protocol was coded by two raters who had undergone extensive training in the coding system. Coding issues were staffed in regular meetings with all members of the clinical coding team to establish greater clarity regarding any ambiguous scenarios. All clarifications to coding procedures were maintained in a written log that served as an addendum to the code book over the project period.

Validation and final coding. Once double-coded, information from studies and protocols was entered into an application that compared all entries for discrepancies across raters. When the two raters agreed, these results were written automatically to a final record, and when raters disagreed, the problematic field was flagged as a discrepancy for an expert reviewer (i.e., Bruce F. Chorpita), who was expected to resolve the discrepancy through a third coding of the relevant study or protocol code. Also, for the official coded record, all fields were given a final inspection for accuracy by an expert reviewer and were subjected to multiple data validation utilities to search for outliers or other offending values.

Winning treatment groups. Study groups were entered into the data set for analysis only if they qualified as a “winning” treatment group, which was defined as a group in which an active, nonpharmacological treatment beat one or more other study groups (e.g., a psychosocial treatment group, medication, a combined psychosocial and medication group, placebo, wait-list, no-treatment, or other control group) in a randomized trial on the primary outcome measure in the target symptom domain (e.g., the primary depression measure in a study of depression). Combined treatments were excluded from the definition of winning treatment groups, because these did not allow clear inferences about whether the psychosocial portion of the combined protocol contributed to the observed group difference. According to our criteria, 279 of the 671 treatment groups (testing 615 different treatments) produced winning treatments in these studies. Reliability of the classification of winning status to a study group was done through a random sample of 10% of the 671 groups and demonstrated excellent reliability ($r_s = .95$ for assignment of treatment rank within study).

Codes for Study Variables

Study codes relevant to this review include those that pertain to age, gender, ethnicity, and problem area (e.g., anxiety, depression). These four variables were selected and employed as context variables (meaning they had the potential to illustrate interaction effects on the patterns of practice frequency) from a theoretically rather large set of potential variables (e.g., nationality, parental marital status, geographic location, treatment setting, treatment format). On the one hand, we restricted the number of study variables to four to limit the exponential increase in computational complexity that arises from increasingly higher order interactions. In addition, we selected variables that in our experience are relatively commonly reported in the literature and that historically

have been used to test moderator effects on treatment outcome or that have been given special emphasis in national practice policy (e.g., HHS, 1999).

For all demographic information, priority was given to information pertaining to specific study groups (e.g., when age and gender was reported separately by groups). When information was not available at the study group level, it was extrapolated from the study level (e.g., if the study level described inclusion of 4- to 7-year-olds, then 4- to 7-year-olds were assumed to be present in each study group). To estimate reliability, we sampled approximately 20% of study groups for the study variables (see Table 1), which are defined below. When codes were not frequent enough in the 20% random sample to allow estimation of reliability, estimates were made using the agreement between the initial two raters prior to validation (a conservative estimate that consistently underestimated the reliability of the validated record when both were available). If a study group had no reliable problem code (i.e., the only problem for which participants were selected for the study could not be reliably coded with a κ of at least .65), the study

Table 1
Overview of Study Variables

Variable	Frequency ^a	κ^b
Age		
0–3 years	37	1.00
4–7 years	96	1.00
8–11 years	122	.86
12–15 years	133	.86
16–19 years	61	.90
Not reported	20	.82
Gender		
Female	177	1.00
Male	197	1.00
Not reported	25	1.00
Ethnicity		
American Indian or Alaska Native	3	1.00
Asian	8	1.00
Black or African American	51	.95
Caucasian	76	.92
Hispanic or Latino/a	26	1.00
Multiethnic	6	1.00
Not reported	145	.93
Population problem area		
Aggression	48	.95
Anger	12	.82
Anxiety	56	.90
Attention	10	1.00
Autism	7	1.00
Avoidance	4	.66 ^c
Depressed mood	24	1.00
Hyperactivity	21	1.00
Justice involvement	12	1.00
Oppositional	45	1.00
Phobia	35	1.00
School refusal/truancy	6	.75 ^c
Shyness	8	.66 ^c
Substance use	6	.71 ^c
Traumatic stress	11	1.00
Willful misconduct, delinquency	39	1.00

^a $N = 232$. ^b $n = 58$. ^c Base rates for these codes were too low to allow estimation in the 20% random sample, so reliability was estimated with the agreement between initial raters (prior to validation) across all winning study groups.

group was dropped from the analysis. This resulted in a final sample of 232 winning study groups.

Each study was examined in aggregate rather than at the level of effects for each variable category or combination of variable categories. Thus, findings from a study that included 8-year-old participants were assumed to be relevant to 8-year-olds, even though 8-year-olds may have represented a minority of the participants (e.g., a study with an age range of 7–15 years). Similarly, in a study of separation anxiety, social anxiety, and generalized anxiety, in which the treatment group beat the wait list, we assumed that the treatment findings applied to all groups included in the study (unless explicitly noted otherwise), although it is theoretically possible that the treatment did not work for one of those three disorders. We see this as a necessary limitation of the literature to some extent, in that results are rarely reported separately by groups (often due to power constraints). For this article, then, we chose to maintain the standard convention in the literature of assuming that the results are likely to apply to those groups that are adequately represented.

Age. Age of participants was coded as the maximum and minimum age reported in each study group or study. When only means and standard deviations were reported, the range was estimated at the mean plus or minus 1.5 standard deviations. When only the mean or no information was provided, age was imputed from grade level. If those data were also missing, age was coded as not reported.

Gender. Gender of participants was coded as whether the study reported any presence of boys or girls. Thus, if a study included at least one boy, it was coded as including boys, and if it included at least one girl, it was coded as including girls. The decision to reduce this code to a binary field for boys and girls was based on the observation that many studies did not report the sample size separately for boys and girls, and among those that did, those with at least one boy (or girl) almost always had at least 30% of its sample as boys (or girls). Thus, the indication that at least one of a particular gender group was present generally indicated adequate representation of that group. When no information was provided, gender was coded as not reported.

Ethnicity. Participant ethnicity was similarly coded using the “at least one” strategy, given that a large number of studies reported ethnic group membership without providing specific numbers or percentages for each group. Ethnicity was coded using the U.S. Census definitions for major groups. When no information was provided, ethnicity was coded as not reported.

Problem area. The nature of the target problem experienced by study participants was coded using a checklist of 25 different problem areas, plus the ability to write in up to three other entries that did not fit the checklist. A *target problem* was defined as that which the study explicitly targeted with the protocol and for which outcomes were measured. Comorbidity among problem areas was coded in all instances, but it was included in analyses only when protocols targeted multiple problem areas simultaneously (e.g., hyperactivity and inattention).

If a problem area was coded as applicable to at least four winning treatment groups, then the problem area code was represented in the potential list of study variables for analysis (autism spectrum and externalizing not otherwise specified were dropped for this reason). Although not technically a population problem per se, juvenile justice involvement (e.g., treatment delivery in a

correctional setting) was coded as its own problem area, given that (a) some studies indicating juvenile justice involvement were not explicit about the population problem and (b) youths in juvenile justice settings could potentially have problems that factor differently from similar population problem area codes (e.g., willful misconduct/delinquency, aggression). This resulted in a list of 23 problem area codes. Using the procedure outlined above, we dropped four of these codes (peer interaction, peer involvement, runaway, and self-control) due to low reliability estimates, leaving 19 problem areas. Three final codes (academic achievement, learning disability/underachievement, and school involvement) were dropped because they represented problem areas that were not directly targeted in the study sampling process. A final set of 16 codes was entered into the final analyses (see Table 1).

Codes for Practice Elements

The protocol from each winning study group was coded for its specific content by two raters regarding the presence or absence of 58 practice elements. We defined a *practice element* as a discrete clinical technique or strategy (e.g., time out, relaxation) used as part of a larger intervention plan (e.g., a manualized treatment program for childhood anxiety; see Chorpita et al., 2005, for an extensive definition and for the development of the initial pool of 58 codes). The coding procedure allowed raters to write in other entries, three of which were common enough to warrant inclusion in analysis (physical exercise, psychoeducational–teacher, and personal safety skills), resulting in an expanded pool of 61 codes. We chose to code treatment operations as opposed to nonspecific effects primarily due to the availability of the descriptions of such variables in the protocol sources. For example, fewer than 1% of the protocols that we coded contained descriptions of supportive listening or relationship/alliance building, even when we explicitly coded for them (see below).

As with the codes for study variables, an expert rater performed a final validation and review of all codes. Coding was performed on the best available description of the treatment procedures, which in the majority of cases was the description provided in the text of a research study. When the actual manuals were available, these were the first choice for coding; however, this occurred for only 31 cases, given that commercially available manuals were often adaptations or later editions of the protocols tested rather than the actual protocols themselves and were thus inadmissible for coding. To be conservative, we coded only manuals whose use was clearly described as the sole and unmodified source of the tested protocol and that was available to us either commercially or through direct requests to the treatment developers.

When practice elements among the 61 codes occurred fewer than three times across all winning study groups, those codes were excluded from the final analyses, given the sensitivity of the kappa statistic to base rate extremes. Fourteen low base rate codes—case management, cultural training, emotional processing, eye movement/tapping, free association, hypnosis, line of sight supervision, mentoring, milieu therapy, mindfulness, motivational interviewing, relationship building, supportive listening, and twelve step programming—were removed for this reason.

The resulting 47 codes were then subjected to reliability analysis, based on a random sample of 20% of the 232 groups (see Table 2). As with the analysis of the study variables, when codes

Table 2
Overview of Practice Element Codes

Code	Frequency ^a	κ^b
Activity scheduling	15	.74
Assertiveness training	23	.79
Attending	19	.85
Behavioral contracting	27	.84
Biofeedback, neurofeedback	3	1.00
Cognitive	89	.85
Commands	34	.84
Communication skills	50	.73
Differential reinforcement	33	1.00
Educational support	15	1.00
Exposure	81	1.00
Family engagement	10	.75 ^c
Family therapy	18	1.00
Goal setting	46	.84
Guided imagery	17	1.00
Insight building	14	1.00
Maintenance/relapse prevention	53	.68
Marital therapy	8	1.00
Modeling	81	.80
Monitoring	38	.93
Natural and logical consequences	19	.66 ^c
Parent coping	18	.72 ^c
Physical exercise	4	1.00
Praise	62	.77
Problem solving	75	.83
Psychoeducational-child	60	.88
Psychoeducational-parent	65	.90
Relaxation	66	.95
Response cost	28	.92
Response prevention	4	.66 ^c
Self-monitoring	52	.81
Self-reward/self-praise	37	1.00
Self-verbalization	4	1.00
Social skills training	45	.80
Stimulus control or antecedent management	26	.74
Talent or skill building	15	1.00
Tangible rewards	66	.84
Therapist praise/rewards	60	.82
Time out	43	.92
Discrete trial training	3	1.00
Personal safety skills	3	1.00

^a $N = 232$. ^b $n = 54$. ^c Base rates for these codes were too low to allow estimation in the 20% random sample, so reliability was estimated with the agreement between initial raters (prior to validation) across all winning study groups.

were not frequent enough in the 20% random sample to allow estimation of reliability, estimates were made using the agreement between the initial two raters prior to validation. Six codes with poor reliability ($\kappa < .65$)—catharsis, crisis management, interpretation, parent coping, play therapy, and psychoeducational-teacher—were dropped at this stage. No clear pattern emerged to suggest a reason for the low reliability, although all of the dropped codes were endorsed less often than the average of the final 41 codes listed in Table 2.

Analytical Approach

The aim of this article was to apply the DMM to a particular set of studies to determine the resulting profiles and structure in the chosen literature (Chorpita et al., 2005), and interested readers are

also referred to the related literature on data mining in information sciences (Brodley, Lane, & Stough, 1999). The data reduction approach was a variant of an interaction-detector algorithm, using intraclass correlation coefficients (ICCs) as the pattern similarity index. We chose the ICC for the following reasons: (a) It is grounded in analysis of variance models similar to methods used in evaluating therapeutic practices relative to third variables (e.g., Trijsburg et al., 2002); (b) it has a known standard error and a convenient scaling with an upper limit of 1 (Webster, 1952); and (c) it is sensitive to aspects of both elevation and scatter, albeit imprecisely (Cronbach & Gleser, 1953; Webster, 1952). Although the ICC is imprecise about the exact type of similarity (profile elevation, scatter, or shape), our present application was intended to test only predictions about gross differences among practice profiles for different context variables.

A high ICC value between different categories of study variables (e.g., the categories boy and girl from the gender variable) thus suggests that the pattern of average frequency of practice elements for those categories is similar. Within each variable, the pair of categories with the highest ICC meeting criterion was merged, and then ICCs were recalculated to determine whether more merges were possible among the newly formed categories. This process repeated until no more merges were possible. Put simply, if girls and boys both received winning treatments that were characterized by similar proportions of practice elements A, B, and C, then the categories girls and boys would merge and would be unable to represent a branching in the tree. Alternatively, if winning treatments delivered to children and adolescent samples were characterized by different proportions of practice elements X, Y, and Z, then those categories of the age variable would not merge and would be available as a potential branch in the tree. Although we used an empirical approach for detecting these differences, these patterns are generally intuitive and quite apparent in visual displays (e.g., in our experience, even untrained judges can discriminate a graph that has mostly exposure and cognitive from a graph that has mostly rewards and time out when presented side by side).

Merged variables were then subject to analysis for producing optimal splits in the tree. Because of large amounts of missing data in the literature (e.g., more than 50% of studies did not report ethnicity), we applied a strict criterion of interpretability for assigning splits. Specifically, any splits producing a node whose sole membership was not reported were considered ineligible (e.g., a split involving boys or girls as one node and gender not reported as the other), because they provided insufficient guidance for interpreting the literature. Among the eligible splits, the variable that was maximally informative was then selected for branching, as defined by having the lowest average pairwise ICC value among its categories. For example, a problem variable reduced to three categories (anxiety, depressed mood, and hyperactivity) might produce three practice element profiles that are more different (less correlated) from each other than two profiles are from each other that were created by an age variable reduced to two categories (child and adolescent). In such a case, problem would be selected as the branch node.

This procedure is then repeated recursively within each resulting branch node, using all remaining study variables. The final point in this process produces a terminal node that defines a unique pattern of average practice element frequency. This node describes the

successful clinical protocols whose component practices distinctly characterize a given area of the literature. Multiple solutions are collectively examined to inform a final distillation tree, for which practice element profiles are presented.

Results

Initial Distillation Trees

Approaching this task in a manner similar to exploratory factor analysis, we chose multiple criterion values by which to merge variable categories (cf. factor extraction), in an attempt to derive four initial trees that varied in their level of empirical differentiation (ICC $p_{\text{merge}} = 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}$, respectively). As expected, decreasing the merge criterion value produced a greater number of nodes for each tree (9, 12, 17, and 22 nodes for the four solutions, respectively). Despite the increasing complexity, the problem study variable was the first variable to split for all solutions (average pairwise ICCs ranged from .16 to .20). Problem nodes ranged in number from 5 to 10, and first-order nodes were present and unchanging for autism, depressed mood, and substance use in all solutions. In the least differentiated 9-node ($p_{\text{merge}} = 10^{-2}$) solution, an anxiety/avoidance/traumatic stress node emerged, from which four study groups split off in the 17-node ($p_{\text{merge}} = 10^{-6}$) and higher solutions to form an avoidance node, and 11 study groups split off in the most differentiated 22-node ($p_{\text{merge}} = 10^{-8}$) solution to form a traumatic stress node. Externalizing problems also differentiated across solutions, forming a large single node in the least differentiated 9-node ($p_{\text{merge}} = 10^{-2}$) solution (aggression, anger, attention, conduct, hyperactivity, justice involvement, oppositional, school refusal/truancy) from which six study groups split off in the 12-node ($p_{\text{merge}} = 10^{-4}$) and higher solutions to form a school refusal/truancy node. In the most differentiated 22-node ($p_{\text{merge}} = 10^{-8}$) solution, the remaining externalizing categories split into three large nodes representing attention deficit/hyperactivity (ADH), oppositional/aggressive, and delinquent category groups. Because the 22-node ($p_{\text{merge}} = 10^{-8}$) solution produced meaningful and interpretable groupings in the problem variable, we chose to retain these for the final model (see Figure 1). Each of these main areas is discussed in turn.

Autism. Autism study groups produced an age split in three of the four solutions, that is, the 12-node ($p_{\text{merge}} = 10^{-4}$) and higher. The split always involved one study group that included participants in the 12–15 years age range (ICC = .53). The other category was a duplicate of the parent node. These results suggest that the only study of treatments for autism to include children ages 12–15 years was characterized by a protocol that is empirically distinct from the aggregate of all the studies in terms of its component practice elements. These results were retained for the final distillation tree (e.g., see Figure 1), and we applied a convention of renaming a duplicate subnode (ages 0–11 years) the same as its parent node (all autism) in the final tree to aid interpretation.

Anxiety. In the most differentiated 22-node ($p_{\text{merge}} = 10^{-8}$) solution, this node produced a split for ethnicity such that five winning study groups with Asian, Hispanic, and multiethnic participants emerged with a distinct practice profile from the overall average for anxiety (ICC = .56). Similar to the results for autism, the patterns suggested a special case for a subgroup of children

within a given problem area. We therefore applied the same convention as above to rename the duplicate node in this split, and this pattern of results was retained for the final tree. In the less differentiated 17-node ($p_{\text{merge}} = 10^{-6}$) and lower solutions that merged traumatic stress and anxiety into a single node, three study groups with Asian participants showed a distinct profile, but because the composite anxiety and traumatic stress node was of a fundamentally different nature, the Asian-only category was not maintained in the final tree.

Avoidance. This category split from anxiety in the 17-node ($p_{\text{merge}} = 10^{-6}$) and higher solutions. In no case did avoidance produce higher order splits. It was included as a single terminal node in the final tree.

Traumatic stress. This category split from anxiety in a most differentiated 22-node ($p_{\text{merge}} = 10^{-8}$) solution only and produced a second-order split for age (ICC = .70), suggesting a special case for a single study group with participants in the 0–3 years age range. The other terminal node represented a duplicate of the parent node and was renamed accordingly.

Depressed mood. Depressed mood displayed a second-order split on ethnicity in all analyses. In the 17-node ($p_{\text{merge}} = 10^{-6}$) and higher solutions, ethnicity split into four categories: (a) Black or African American, (b) Hispanic or Latino/a, (c) Multiethnic, and (d) all other (see Figure 1). In the less differentiated 12-node ($p_{\text{merge}} = 10^{-4}$) and lower solutions, ethnicity split into two categories, namely (a) Black or African American or Multiethnic and (b) all depressed mood, a duplicate of the parent node. Interestingly, in the least differentiated solution, depressed mood showed a third-order split for age (ICC = .24) such that a study group with 4- to 7-year-olds emerged as distinct from the entire group. This split was pruned in the more differentiated solutions because the age not reported category split into its own group, which failed to satisfy the interpretability criterion for maintaining splits. The four-category ethnicity split was retained for the final tree.

School refusal and truancy. This category split off from all externalizing categories under the 17-node ($p_{\text{merge}} = 10^{-6}$) and higher solutions, yielding no further splits. It was thus retained as a terminal node for the final tree.

Attention deficit/hyperactivity (ADH). This node split from the externalizing node in the most differentiated 22-node ($p_{\text{merge}} = 10^{-8}$) solution only, with 22 study groups showing a distinct profile for two problem categories. It was retained as a terminal node for the final tree.

Oppositional/aggressive. This node split from the externalizing node in the highest branching solution only, with 68 study groups showing a distinct profile for three problem categories. It was retained as a terminal node for the final tree.

Delinquent. This node split from the externalizing node in the highest branching solution only, with 39 study groups showing a distinct profile for two problem categories. It was retained as a terminal node for the final tree.

Substance use. This category yielded a terminal node in the first two trees and produced an identical binary split for age in the 17-node ($p_{\text{merge}} = 10^{-6}$) and higher solutions. In those analyses, the results suggested that the subset of winning study groups including participants from ages 4 to 11 years yielded a distinct practice element profile. This split was retained for the final tree and had an ICC of .61.

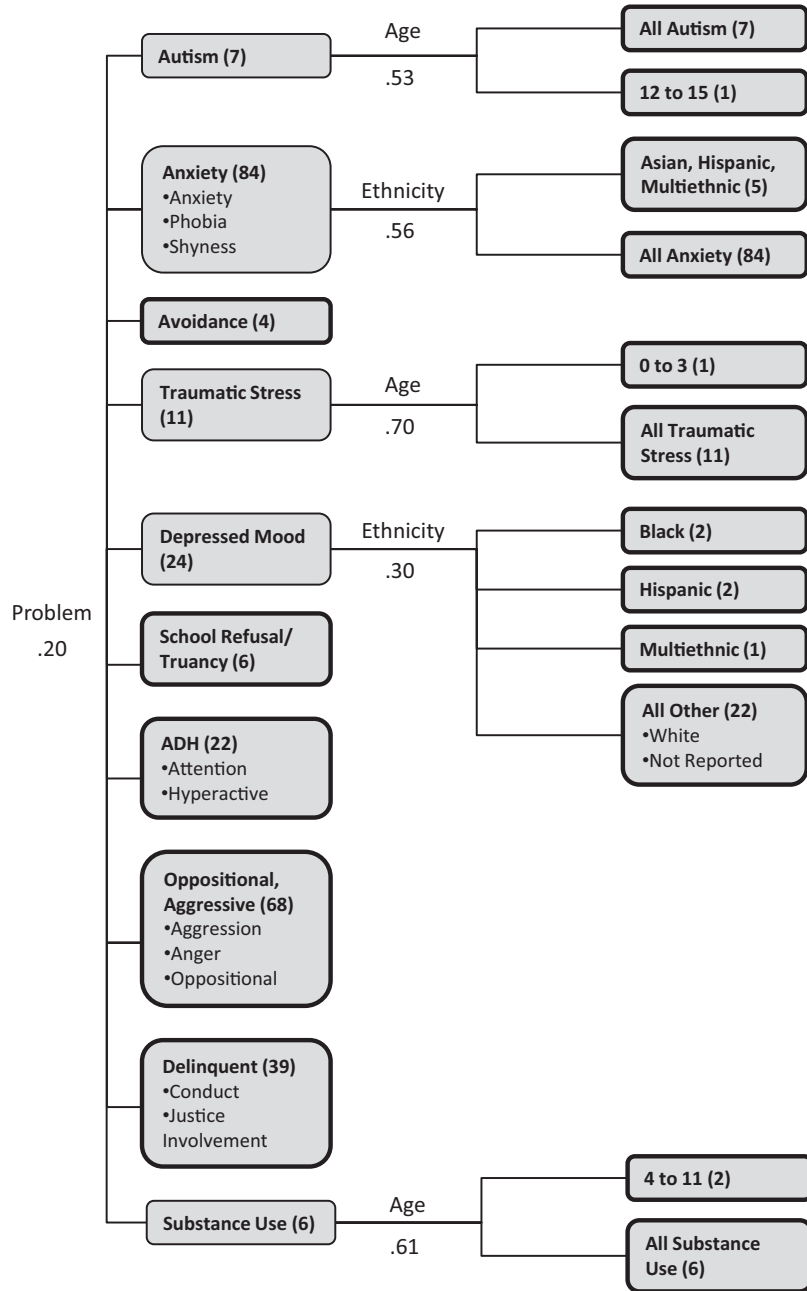


Figure 1. Final working distillation tree. Bold headings represent node name. When constituent categories are ambiguous, these are listed below the heading. The number of study groups in each node is shown in parentheses. Bold outlines represent terminal nodes. Numbers by branches represent the average pairwise intraclass correlation coefficients among resulting nodes. ADH = attention deficit/hyperactivity.

Practice Element Profiles for Selected Nodes

Autism. The terminal nodes for autism appear in Figure 2. For autism overall, the most common practices involved teaching skills to address the core communication and social deficits of the disorder through modeling, goal setting, parent education, and various forms of reinforcement (i.e., communication skills, modeling, social skills training, and goal setting). The greatest differences between nodes showed that the one study group including 12- to

15-year-olds used the less common strategies of psychoeducation and building problem-solving skills and insight with youths.

Anxiety, avoidance, and traumatic stress. Figure 3 shows the practice element profiles for all nodes in the final solution that originally emanated from the anxiety/avoidance/traumatic stress node in the least differentiated solution. The most common practices for anxiety included exposure, relaxation, cognitive, modeling, and psychoeducational-child. Within the subset of five dis-

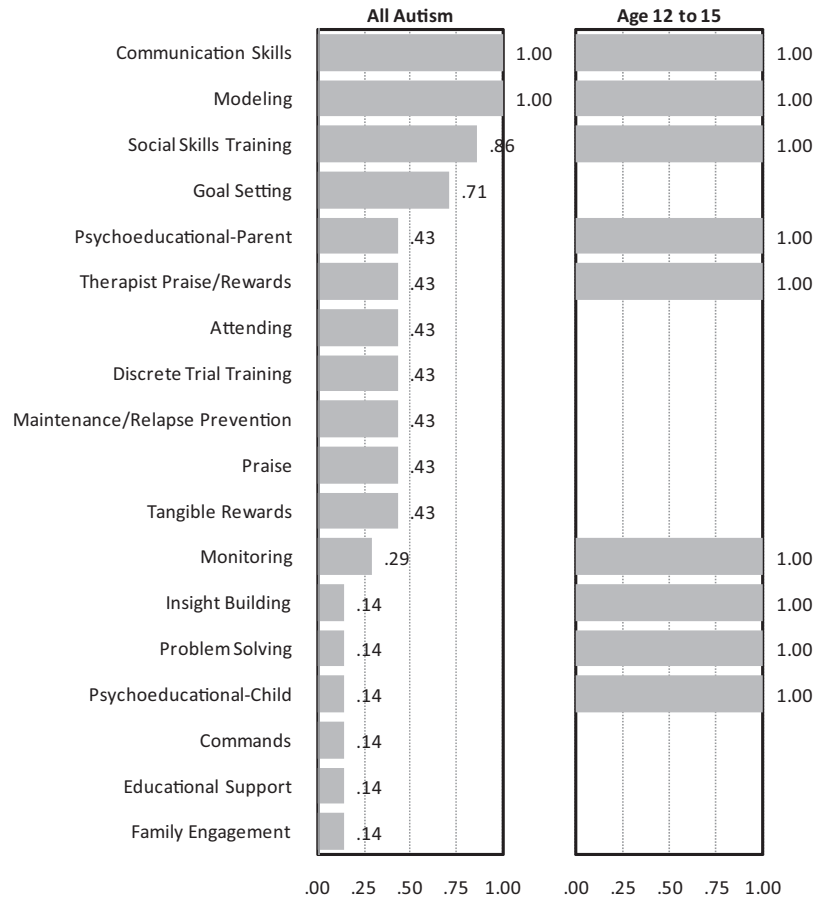


Figure 2. Practice element profiles for selected nodes within the autism branch.

tinct study groups reporting Asian, Hispanic, or multiethnic participants, the profile differed in its relatively greater emphasis on parent coping, communication skills, behavioral contracting, tangible rewards, and psychoeducational–parent.

The avoidance node appeared to capture those studies for which phobic avoidance was targeted explicitly (these four study groups are a subset of the original 94 anxiety/avoidance/traumatic stress groups from the least differentiated solution). The profile showed the uniform use of exposure paired with therapist praise/rewards, along with modeling, in one of the four protocols.

The all traumatic stress node was similar to the all anxiety node; however, it included two skills (albeit with low frequency) that never appeared in the 84 study groups targeting anxiety only: personal safety skills (appearing in 3 of 11 study groups) and insight building (appearing in 1 of 11). Also notable were (a) a greater emphasis on cognitive and psychoeducational–child relative to all anxiety study groups and (b) the absence of therapist praise/rewards and training the child in self-reward. The profile for the single winning study group with participants from ages 0 to 3 years excluded common practices of maintenance/relapse prevention, modeling, communication skills, personal safety skills, and self-monitoring but was otherwise similar to the parent node for traumatic stress in general.

Depressed mood. Unlike most other areas, the depressed mood node did not yield a terminal node that was a duplicate of

itself. Thus, for descriptive purposes, the node for all depressed mood study groups is presented along with all terminal nodes from the final solution (see Figure 4). The most common practices for depressed mood overall were cognitive, psychoeducational–child, maintenance/relapse prevention, activity scheduling, problem solving, and self-monitoring. The subset of study groups including Black participants was characterized by a much smaller number of practices, including cognitive, communication skills, and a higher frequency of family therapy. The subset of study groups including Hispanic participants was roughly similar to that for all depression but with a higher frequency of psychoeducational–parent and parent coping and the notable absence of goal setting and self-reward/self-praise. One of the two study groups also used insight building as a component of interpersonal psychotherapy (Mufson, Weissman, Moreau, and Garfinkel, 1999), whereas the other also used exposure for the treatment of trauma in Latino immigrants (Kataoka et al., 2003), both of which were uncommon in the all depression node. The single study group with multiethnic participants used only the single practice of cognitive. Finally, the node for participants whose ethnicity was White or not reported yielded a set of study groups that overlapped highly with the overall depressed mood node; hence, the practice element profiles were almost identical, with less use of parent coping and exposure being notable.

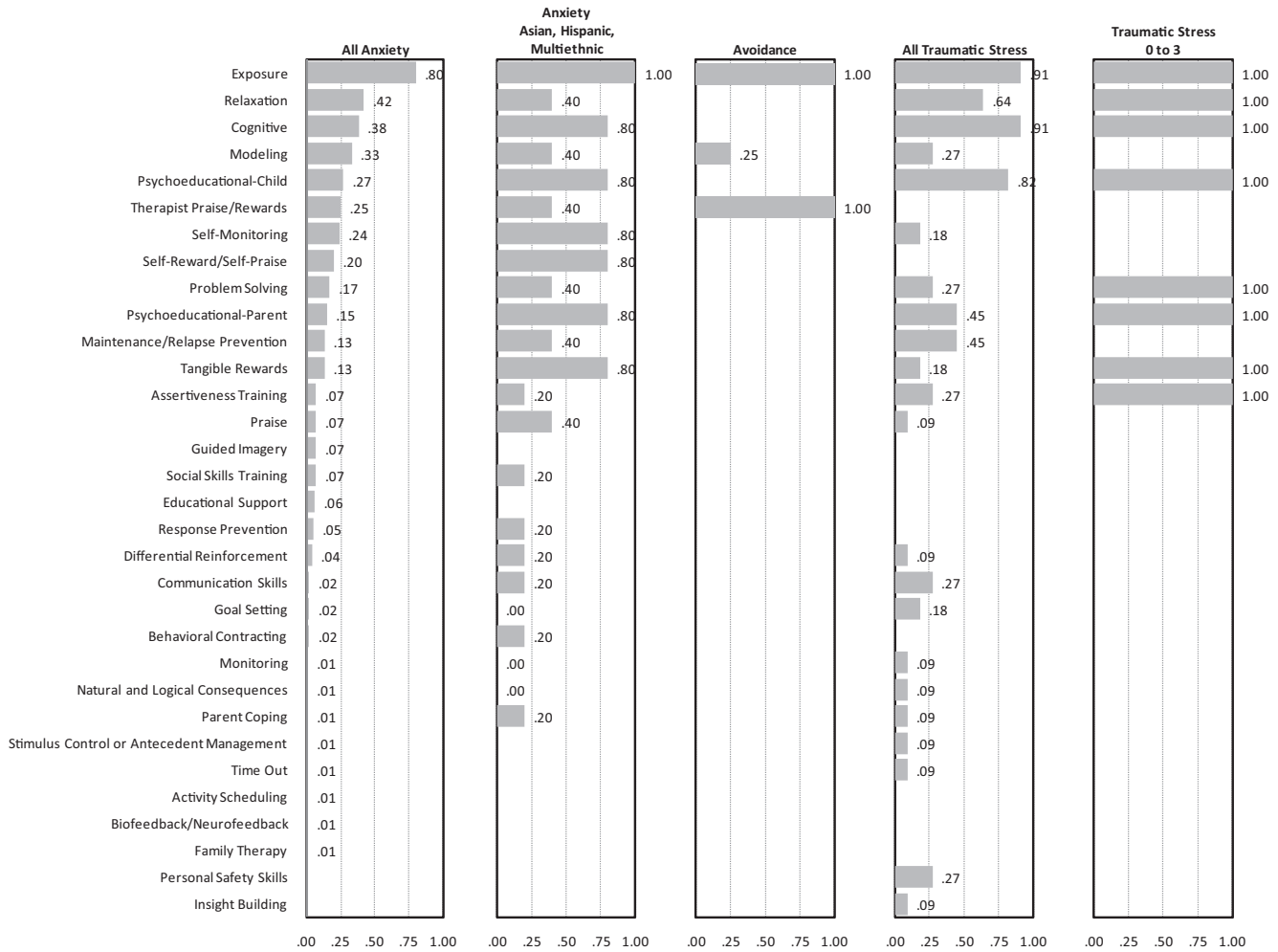


Figure 3. Practice element profiles for selected nodes within the anxiety and stress branches.

Externalizing nodes. Figure 5 shows the practice element profiles for all nodes in the final solution that originally emanated from the externalizing node in the least differentiated solution. Four nodes, each with rather distinct practices, were produced. The most common practices for oppositional/aggressive were praise, time out, tangible rewards, commands, problem solving, and differential reinforcement. In the delinquent node, the most common practices were problem solving, tangible rewards, praise, cognitive, monitoring, response cost, and social skills training. Relative to the oppositional/aggressive node, roughly twice as many of the most common practices targeted child skills as opposed to caretaker skills. Also notable was the greater emphasis on family issues, including practices of family engagement, family therapy, and marital therapy for the caretakers.

In the ADH node, the emphasis on the parent skills of praise, time out, rewards, and psychoeducational–parent was similar to that of the oppositional/aggressive node. A notable difference from other externalizing nodes was the complete absence of cognitive, behavioral contracting, attending, self-monitoring, and psychoeducational–child. Physical exercise was also relatively more common in this node than in others, and self-verbalization

and biofeedback were unique to this node. Finally, the school refusal/truancy node showed a large number of practices with only a modest level of use across studies. This suggested that the protocols from study groups targeting school refusal were likely heterogeneous in nature. The most common practice elements in this node were goal setting and self-monitoring.

Substance use. Figure 6 shows the overall practices for substance use and for the unique node associated with participants ages 4 to 11 years. The most common practices overall included family therapy, communication skills, assertiveness training, family engagement, modeling, self-monitoring, and stimulus control or antecedent management. The protocols tested in children ages 4 to 11 years, on the other hand, both focused on family therapy, with one also using communication skills and the other psychoeducational–parent.

Discussion

What the Results Tell Us

This is the first broad aggregate summary of the components of successful treatments tested in randomized trials for children,

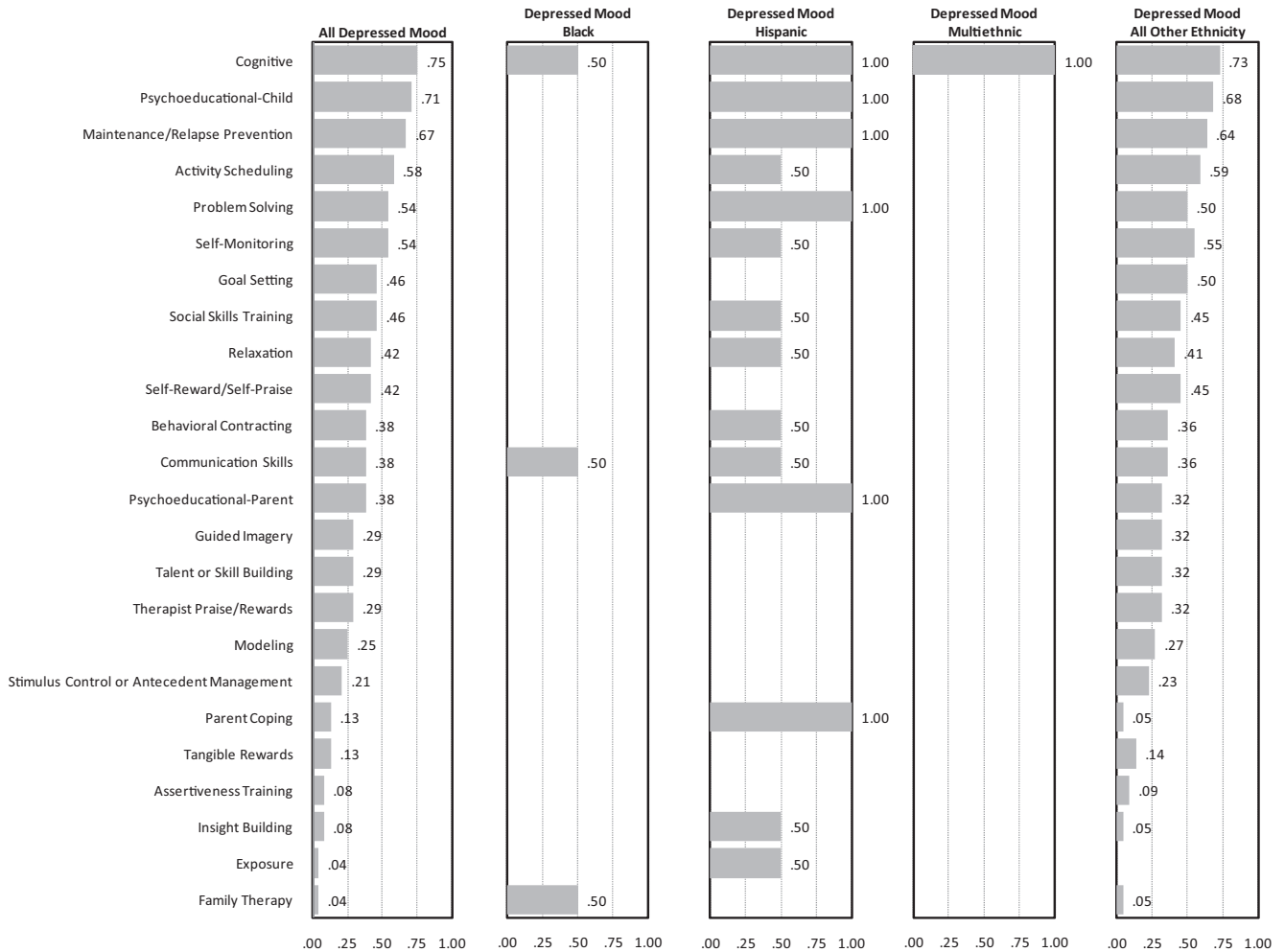


Figure 4. Practice element profiles for selected nodes within the depressed mood branch.

organized from the practices upward rather than from rationally selected groupings. The patterns show that treatments are characterized by a large number of practice elements that vary considerably depending on different variables of interest. Child problem area was the primary factor by which the component practices of successful treatments were organized, which lends some empirical support to the traditional “disorder-driven” approach to other reviews. One clinical implication of the resulting practice element patterns is that they can serve as a guidepost for either (a) the selection of an evidence-based protocol that is potentially “most representative” of the literature in which it resides (i.e., which manual looks most like its node) or (b) the ad hoc design of a treatment (e.g., modification of usual care) through selection of practice elements that appear relevant to a given group.

The organization of the resulting tree was roughly characterized by two patterns: “mature splits,” which were based on a large number of nonoverlapping observations, and “special cases,” which represented a small subset of study groups with a unique profile. Most of the problem areas in our final tree represented mature splits (with the exception of avoidance, which was a complete subset of the anxiety node), whereas the

subsequent splits were mainly special case scenarios (e.g., a single study group for autism with older participants). Problem area was somewhat more predisposed to mature splits, given that study groups tended to have membership in a limited number of categories, whereas for the other study variables (e.g., ethnicity, gender) most study groups fell into a large proportion of the categories, making nonoverlapping splits difficult. In some of the intermediate solutions, however, age appeared as a mature split (i.e., older and younger) within some nodes, which fits with an intuitive sense of the literature. However, these findings were not robust across solutions, and these patterns of differences (e.g., between caregiver and individual-based techniques for externalizing) were ultimately better explained by problem area (i.e., oppositional vs. delinquent better than younger vs. older). The tendency for the literature to develop mature splits for problem area may be a function of a research tradition organized around psychopathological groups more than anything else.

Special cases means more than just that particular groups were represented in the literature; it means that when they were represented, the treatments were characterized by different practice

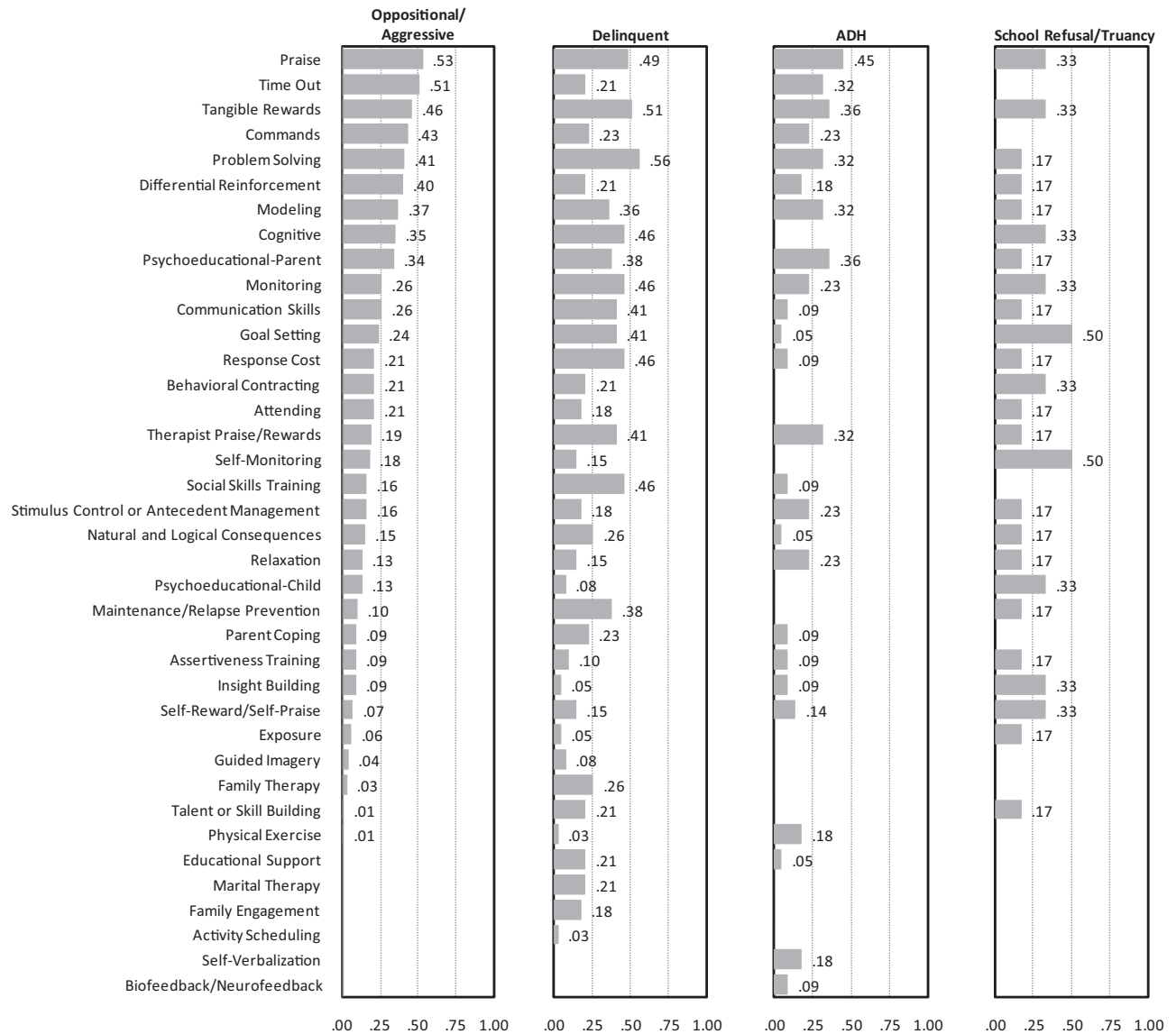


Figure 5. Practice element profiles for selected nodes within the externalizing branches. ADH = attention deficit/hyperactivity.

elements than in the larger literature in which that study was nested. Thus, a second clinical implication of these findings is that special cases should play a greater role in any treatment selection heuristic. For example, when considering the research in planning treatment for a 13-year-old with autism, one should look not only to the larger evidence base of seven randomized trials but also to the single special case study whose population (a) included members that were similar and (b) whose treatment practices were distinct from the larger autism literature. Likewise, given our special case finding that training parents to praise was more common in studies of anxiety treatments with Asian children (40% vs. 7% for all anxiety), one should consider whether to include praise when working with Asian families, to the extent that those five studies might be more germane than the larger anxiety liter-

ature. In that sense, the DMM provides a better definition of the hierarchical sets of the literature from which the clinician should generalize than other reviews might.

As mentioned elsewhere (Chorpita & Daleiden, 2007; Chorpita et al., 2005), these results have at least two clear implications for researchers: (a) suggesting new avenues for treatment design by highlighting the aggregate characteristics of practice components across the literature and (b) identifying areas of the literature that are comparatively underdeveloped. Regarding the first implication, researchers seeking to modify and test treatments can systematically consider the common practices in a given domain and ask questions about the performance of new combinations or subsets of those elements (e.g., testing a two-component vs. a multicomponent depression protocol). The second implication

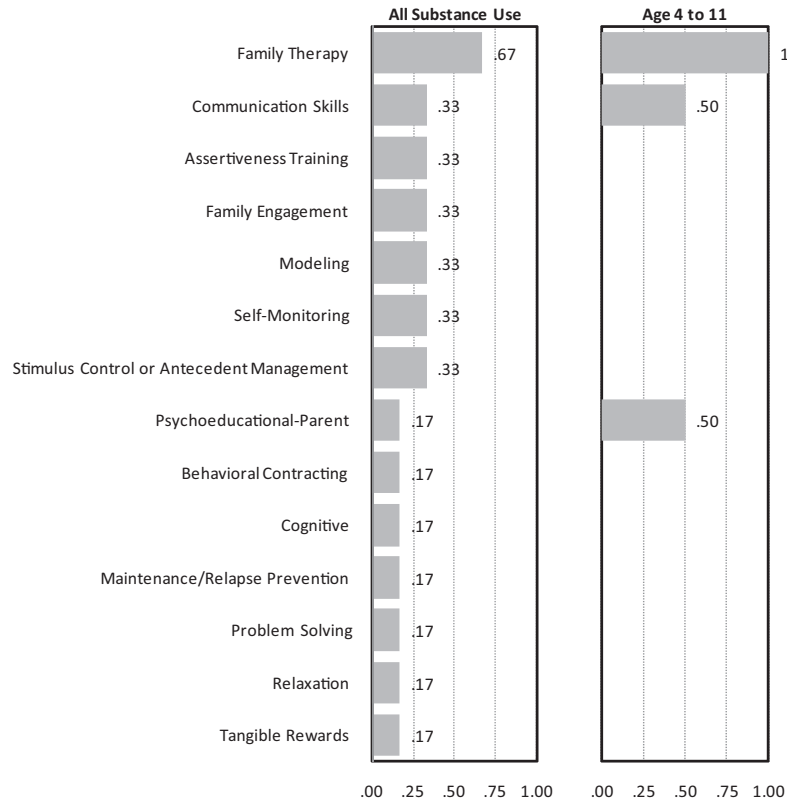


Figure 6. Practice element profiles for selected nodes within the substance use branch. ADH = attention deficit/hyperactivity.

highlights how any special case scenario might be tested relative to the parent node regarding the importance of practices unique to the parent node (e.g., how important is praise to Asian children with anxiety?). It also suggests pockets of the literature that may simply have too few studies (e.g., depression in Hispanic youths).

What the Results Do Not Tell Us

Although we have illustrated how one could generate hypotheses for prescriptive heuristics from these findings, the findings are primarily descriptive in nature. Thus, although we recommend that the clinician consider praise when treating Asian children with anxiety, we cannot conclude that praise is more efficacious with Asian children than not including praise. That inference is possible only from a comparative trial within that population. As with all descriptive reviews, and the literature more generally, much more can be said about what worked than about why something worked.

More generally, the practice element level of analysis can lend itself to misconceptions about necessity, sufficiency, and efficacy of individual elements. Although in limited circumstances one can claim that certain practice elements appear sufficient for outcomes (e.g., cognitive bibliotherapy for multiethnic youths with depressed mood; Ackerson et al., 1998), this model (and the literature on which it is based) is not organized to examine such claims. Inferences about necessity are even more challenging; thus, particular caution is warranted with respect to such terms as *evidence-based practice elements*.

This analysis also cannot speak to the relation of variables outside the model (e.g., treatment setting) to practice element profiles, nor even whether practice elements themselves are the most important part of successful protocols (as opposed to assignment of homework, therapeutic alliance, etc.). Although it does seem plausible that by definition practice elements would differ more greatly across client characteristics than would nonspecific factors, such as warmth and alliance, that variability alone is not a testament to their importance in achieving therapeutic outcomes. Similarly, the focus on practice elements overlooks many other potentially important features of evidence-based practices besides their components, features such as their sequencing, coordination, and guiding supervision infrastructure (Chorpita & Daleiden, 2007). Finally, because coding was performed to identify only the rough structure of practices, some group differences on variables in our models may have been masked. For example, the analysis examined whether rewards were used but did not explicitly discriminate between a sticker chart and a token economy. Overall, then, this review does not tell us what a good treatment should look like as much as what its component practices are likely to be at a given level of specificity.

At times, protocol patterns differed in ways that were unrelated to variables in the model. For example, in the ADH node no practice occurs in even half of the study groups, denoting that many protocols used nonoverlapping sets of elements. In fact, ADH node study groups tested a diversity of approaches, including

self-verbalization (with and without problem solving), biofeedback on its own, physical exercise on its own, or a cluster of elements related to parent management training. With factor analysis, one could see whether different practice element factors emerged within a given node (producing something analogous to “simple structure” within a node); however, the current tree tells us that to choose among those factors is likely a choice that is not uniquely associated with problem area, gender, age, or ethnicity within a terminal node. In other words, there might be four treatment factors for ADHD problems, and the choice of which is best is likely better driven from other considerations available in more traditional meta-analysis (e.g., effect size, replication trials) as well as local individual- or practice-based evidence (Daleiden & Chorpita, 2005). Similarly, designing a treatment ad hoc from a node with less simple structure could produce a protocol unlike anything in the literature (e.g., exercise and time out). Nevertheless, in the face of an unsuccessful treatment plan, these profiles may still offer promising avenues for strategies that could have been overlooked.

Methodological Issues and Limitations

As with other exploratory multivariate techniques, the final analytic results are sensitively dependent on the nature of the data analyzed and the series of strategic decisions made when implementing the analysis. The nature of the current data is affected by the volume and quality of the articles reporting the randomized trial results and the quality and inclusiveness of the codes used to extract information from the articles. Key analytic decisions included selection of a pattern similarity index, selection of criteria for merging and splitting factors, and decisions about pruning.

The volume of available data for analysis (i.e., the overall size of the literature, which ultimately yielded 232 winning study groups) was small for exploratory analysis and represents a fundamental limitation to the study. Further, missing data in the literature was extensive: 8% for gender, 11% for age, and 63% for ethnicity. Only 82 (39.3%) of the study groups had complete data on all three variables. As shown in Table 1, we were able to reliably determine when basic demographic information was not reported, so this is a limitation of the literature itself, not of our approach. Inclusion of splits involving unique not specified nodes (all data missing) yielded trees that were far more elaborate (e.g., 99 total nodes for $p_{\text{merge}} = 10^{-8}$). This means that there may in fact be differences that occurred across groups of studies that are not represented here, because the variables were not documented. The meaning of the undefined nodes could have been any number of the following: a correlation between practice elements and a marker variable for study quality, a cohort effect (e.g., older studies for which reporting standards were more lax), investigator effects (e.g., a single investigator team with three or four studies of the same or similar protocols, with the same incomplete approach for writing participant description), journal effects (e.g., journals with more lax reporting standards tending to accept articles on treatments that are less mainstream). As we attempt to understand the literature more completely, all of these questions are worthy of follow-up analysis in future research.

The analysis also involved selection of a final code set that was based on evidence for the reliability of applying the codes. This strategy increased the likelihood that the results were replicable but decreased the comprehensiveness of the code set for describing

the practice domain. The presence of additional factors that differentiate the literature but were not reliably coded could lead to greater differentiation of the distillation tree. For example, specific codes for obsessions and compulsions were not included, so that even though response prevention was coded as a practice, it was not possible for obsessive-compulsive problems to emerge as a unique node. Likewise, the exclusion of more detailed technique coding within the family therapy or play therapy practice elements may have precluded identification of age-related differences in practice profiles.

With respect to analytic decisions, the selection of the ICC as the pattern similarity index was expected to affect the nature of the observed results, and selection of a different index may yield a somewhat different tree. The ICC is more sensitive to differences in the relative ordering of practice elements, whereas a less correlational measure would be more sensitive to variation in profile elevations. For example, a distance measure might differentiate a profile where a set of 10 practices each emerged in 25% of study groups compared with a profile where the same 10 practices each emerged in 90% of study groups. As illustrated in the results, the present analysis explored a wide range of criterion thresholds for constructing the tree that systematically affected the nature of particular nodes and the overall complexity of the tree. Decreasing the size of the merge criterion beyond the 10^{-8} level would result in increasing differentiation of the problem areas until ultimately each unique code in the original data set becomes its own node of the tree. Finally, data-mining strategies tend to overspecify the initial data so that the final step in data mining involves pruning on the basis of expert judgment. In the present analysis, we exclusively applied pruning rules that were based on the emergence of nodes uniquely defined by missing data. Were these rules not applied, the final tree would have included many more branches, including downstream branches from nodes exclusively defined by missing data.

Taken together, these limitations create a result set of unknown reproducibility. The current analyses are encouraging in that systematically varying the criterion threshold yielded a set of results with numerous points of convergence and meaningful systematic variation that paralleled other expert reviews of child treatments (Weisz et al., 2004). For example, externalizing problems systematically differentiated into the specific nodes of oppositional/aggressive, delinquent, ADHD, and school refusal and truancy. As the literature expands, future analysis of a larger data base with expanded code sets with alternative similarity indexes will help identify true points of convergence.

Summary

This is the first scholarly effort to examine the components of evidence-based practices broadly across the literature, and it represents a new avenue for the mapping of a literature that should continue to become more complex over time. Overall, the literature describes a pattern by which successful treatments that differ in composition organize themselves into clusters that correspond first and foremost to child problem areas. Within problem areas, a variety of special cases exist, suggesting that some child characteristics are associated with a unique pattern of treatment practices. The model overall produced a working map of the literature that quickly points both the clinician and researcher toward larger

collections of relevant study groups and the small number of exceptions that may be germane when considering treatment selection or design. Relative to the current tradition of generalizing from rationally defined problem areas alone, the DMM more clearly shows when the literature may be applicable to a particular subgroup, as well as when that literature has defined a unique pattern of practices. As new studies emerge that push the limits of sample selection, therapist background, and other variables relevant to effectiveness (Schoenwald & Hoagwood, 2001), the DMM can incorporate those new factors and continue to map whether and how the new treatments might look different from the old.

References

- Ackerson, J., Scogin, F., McKendree-Smith, N., & Lyman, R. D. (1998). Cognitive bibliotherapy for mild and moderate adolescent depressive symptomatology. *Journal of Consulting and Clinical Psychology, 66*, 685–690.
- Beutler, L. E. (2002). The dodo bird is extinct. *Clinical Psychology: Science & Practice, 9*, 30–34.
- Brodley, C. E., Lane, T., & Stough, T. M. (1999). Knowledge discovery and data mining. *American Scientist, 87*, 54–61.
- Chambless, D. L., Sanderson, W. C., Shoham, V., Johnson, S. B., Pope, K. S., Crits-Christoph, P., et al. (1996). An update on empirically validated therapies. *Clinical Psychologist, 49*, 5–18.
- Chorpita, B. F., & Daleiden, E. L. (2007). *2007 biennial report: Effective psychosocial interventions for youth with behavioral and emotional needs*. Retrieved from Hawaii Department of Health, Child and Adolescent Mental Health Division: <http://hawaii.gov/health/mental-health/camhd/library/pdf/ebs/ebs012.pdf>
- Chorpita, B. F., Daleiden, E. L., & Weisz, J. R. (2005). Identifying and selecting the common elements of evidence-based interventions: A distillation and matching model. *Mental Health Services Research, 7*, 5–20.
- Clarke, G. N., Rohde, P., Lewinsohn, P. M., Hops, H., & Seeley, J. R. (1999). Cognitive-behavioral treatment of adolescent depression: Efficacy of acute group treatment and booster sessions. *Journal of the American Academy of Child & Adolescent Psychiatry, 38*, 272–279.
- Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*, 456–473.
- Daleiden, E., & Chorpita, B. F. (2005). From data to wisdom: Quality improvement strategies supporting large-scale implementation of evidence-based services. *Child and Adolescent Psychiatric Clinics of North America, 14*, 329–349.
- Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature* (FMHI Publication No. 231). Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network.
- Kataoka, S. H., Stein, B. D., Jaycox, L. H., Wong, M., Escudero, P., Wenli, T., et al. (2003). A school-based mental health program for traumatized Latino immigrant children. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*, 311–318.
- Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63*, 146–159.
- Luborsky, L., Rosenthal, R., Diguier, L., Andrusyna, T. P., Berman, J. S., Levitt, J. T., et al. (2002). The dodo bird verdict is alive and well—mostly. *Clinical Psychology: Science and Practice, 9*, 2–12.
- Mufson, L., Weissman, M. M., Moreau, D., & Garfinkel, R. (1999). Efficacy of interpersonal psychotherapy for depressed adolescents. *Archives of General Psychiatry, 56*, 573–579.
- National Advisory Mental Health Council Workgroup on Child and Adolescent Mental Health Intervention Development and Deployment. (2001). *Blueprint for change: Research on child and adolescent mental health*. Washington, DC: Author.
- PracticeWise. (2005). *Psychosocial and combined treatments coding manual*. Satellite Beach, FL: Author.
- Rounsaville, B. J., & Carroll, K. M. (2002). Commentary on dodo bird revisited: Why aren't we dodos yet? *Clinical Psychology: Science and Practice, 9*, 17–20.
- Schoenwald, S. K., & Hoagwood, K. (2001). Effectiveness, transportability, and dissemination of interventions: What matters when. *Psychiatric Services, 52*, 1190–1197.
- Silverman, W. K., & Hinshaw, S. P. (2008). The second special issue on evidence-based psychosocial treatments for children and adolescents: A 10-year update. *Journal of Clinical Child and Adolescent Psychology, 37*, 1–7.
- Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore: Johns Hopkins University Press.
- Task Force on Promotion and Dissemination of Psychological Procedures. (1995). Training in and dissemination of empirically-validated psychological treatments: Report and recommendations. *Clinical Psychologist, 48*, 3–23.
- Task Force on Psychological Intervention Guidelines. (1995). *Template for developing guidelines: Interventions for mental disorders and psychosocial aspects of physical disorders*. Washington, DC: American Psychological Association.
- Trijsburg, R. W., Frederiks, G. C. F. J., Gorlee, M., Klouwer, E., den Hollander, A. M., & Dulvenvoorden, H. J. (2002). Development of the Comprehensive Psychotherapeutic Interventions Rating Scale (CPIRS). *Psychotherapy Research, 12*, 287–317.
- U.S. Department of Health and Human Services. (1999). *Mental health: A report of the surgeon general*. Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health.
- Webster, H. (1952). A note on profile similarity. *Psychological Bulletin, 49*, 538–539.
- Weisz, J., Hawley, K. M., & Doss, A. J. (2004). Empirically tested psychotherapies for youth internalizing and externalizing problems and disorders. *Child and Adolescent Psychiatric Clinics of North America, 13*, 729–815.
- Weisz, J. R., Weiss, B., Han, S., Granger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin, 117*, 450–468.

Received January 15, 2008

Revision received September 24, 2008

Accepted October 14, 2008 ■