
Article:
Bone, C., Delgadillo, J. and Barkham, M. (2021) A systematic review and meta-analysis of the good-enough level (GEL) literature. Journal of Counseling Psychology, 68 (2). pp. 219-231. ISSN: 0022-0167

https://doi.org/10.1037/cou0000521

A Systematic Review and Meta-Analysis of the Good-Enough Level (GEL) Literature

Article in *Journal of Counseling Psychology* (2020) Pre-Print

Authors: Claire Bone, Jaime Delgadillo, & Michael Barkham

Clinical Psychology Unit, Department of Psychology, University of Sheffield, UK

Acknowledgements: We would like to thank Dr Lewis Hanney for supporting the inter-rater reliability analysis for this study. We would also like to thank the authors who corresponded with us, providing further literature or commentary on their research.

Abstract

Introduction: The "Good-Enough Level" (GEL) model proposes that people respond differentially to psychotherapy, and that the typical curvilinear "dose-response" shape of change may be an artefact of aggregation. We conducted a systematic review and meta-analysis of the GEL literature to examine 1) whether different sub-groups of adults accessing psychotherapy respond to therapy at different rates and 2) whether the shape of change is linear or non-linear.

Method: This review was pre-registered on PROSPERO. Fifteen studies were synthesized (n= 114,123), with 10 included across two meta-analyses (n = 46,921; n = 41,515). Systematic searches took place using Medline, PsycINFO and Scopus databases. A key inclusion criterion was that cases must be stratified by treatment length to examine the GEL.

Results: In support of the GEL, there was no overall association between treatment duration and outcomes (r = -0.24 [95% CI = -0.70, 0.36], p = 0.27). Longer treatments were associated with higher baseline symptom scores (r = 0.15 [95% CI = 0.08, 0.22], p < .001) and slower rates of change. Different shapes of change were also evidenced: curvilinear responses were more often found in shorter treatments, whilst linear shapes were more often found in longer treatments. However, findings varied depending on methodological criteria used.

Conclusion: Although rates of change varied in line with the GEL, most people nonetheless responded within defined boundaries as described in the dose-response literature. We therefore refer to the notion of "*boundaried responsive regulation"* to describe the relationship between treatment duration and outcomes.

Keywords: psychotherapy; outcomes research; good-enough level; dose-response; treatment duration

Public Significance Statement

This review refers to the notion of "boundaried responsive regulation" to describe responses to psychological care. People may respond at different rates and not all follow a curvilinear shape of change, however most will improve within defined boundaries. Overall, this suggests that the duration of therapy should be planned flexibly, in response to client need, yet within boundaries indicated by empirical studies.

A Systematic Review and Meta-Analysis of the Good-Enough Level (GEL) Literature

The duration and cost of psychological care varies considerably across clients. Deciding how long therapy should last and when the outcomes of an individual's treatment have reached a good enough level is therefore a key challenge for clinicians. This question has been a matter of debate in the literature for several decades given its clinical, ethical and economic implications (Harnett, O'Donovan, & Lambert, 2010; Kadera, Lambert, & Andrews, 1996). Two prominent perspectives on the number of sessions required to benefit from therapy include the dose-response (DR) (Howard, Kopta, Krause, & Orlinksy, 1986) and good-enough level (GEL) models (Barkham et al., 1996).

According to the DR model, the relationship between treatment duration (typically measured in sessions) and outcomes is characterized by a negatively accelerating curve, whereby symptomatic improvement mostly occurs in the early stages of treatment and tends to diminish thereafter. Key assumptions of this model are that most people tend to follow this curvilinear response pattern and that the duration or "dose" of treatment *causes* changes to occur, but this effect tends to lessen over time (Howard et al., 1986). Numerous studies over the last 30 years have reported curvilinear DR relationships, as documented in a recent systematic review (Robinson, Delgadillo, & Kellett, 2019). However, there is considerable heterogeneity across these studies regarding the time-point at which treatment gains are observed to diminish, resulting in inconsistent recommendations for an "optimal dose" of treatment. For example, Robinson et al. (2019) reported that optimal doses could vary between 4-54 sessions depending on samples used.

Barkham et al. (1996) pointed out that the DR pattern may partly be a function of aggregating data across different subgroups of cases, some of which complete treatment after only a few sessions and others that have atypically lengthy interventions. The decelerating

shape of change may therefore be a statistical artefact, influenced on the one hand by rapid responders with short treatments, and on the other by gradual and non-responders receiving lengthier treatments. On this basis, treatment duration has been argued to result from *responsive regulation* by clients and clinicians (Stiles, Honos-Webb, & Surko, 1998), where treatment continues until a good-enough level of improvement is attained. According to the GEL perspective, treatment duration is not a determinant of improvement, but rather a function of clients' responsivity to therapy. The probability of improvement would therefore be considered to be either unrelated or negatively related to treatment duration, since non-responders are assumed to have lower probabilities of improvement (Barkham et al., 2006).

Table 1. ***Key Differences between DR and GEL Models***

Dose-Response	**GEL**
Curvilinear response is an average of multiple individual curvilinear responses	Curvilinear response is an artefact of aggregating people, where faster remitters end therapy earlier (the GEL model does not prescribe any particular shape of change)
Rate of change does not vary with total sessions	Rate of change does vary with total sessions
Improvement is associated with total sessions	Improvement is not associated (or negatively) with total sessions
Therapy length determines progress	Progress determines therapy length

A number of studies have found support for the GEL model (for discussions, see Castonguay, Barkham, Lutz, & McAleavey, 2013; Nielsen, Bailey, Nielsen, & Pedersen, 2016). However, unlike the DR literature, no systematic reviews or meta-analyses of the GEL literature have been conducted to date. This means that the distinctive assumptions of the GEL model have not been comprehensively examined across studies. The aim of the present

study therefore was to synthesize the GEL literature using systematic review and meta-analytic methodology. The review was guided by two research questions relating to key assumptions of the GEL model: first, do different sub-groups of adults accessing psychological care respond to treatment at different rates? Second, is the shape of change linear or non-linear?

Method

Protocol Registration

The review protocol was prospectively registered in the PROSPERO database at https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=131840.

Eligibility Criteria and Search Strategy

Table 2 describes the research questions and inclusion and exclusion criteria that guided this study. A systematic search strategy was applied in three databases: Medline, PsycINFO and Scopus. Search terms included variants of: good-enough level, dose-response, treatment duration, rate of change, treatment outcome, responsive regulation and psychotherapy. Search terms were combined using Boolean operators to search within titles, abstracts, keywords or subject headings. No date restrictions were applied. Titles and abstracts were screened by the first author, followed by a full-text review to determine eligibility. Further searches included reverse and forward citations of all selected studies, reference list searches, and email requests for additional recommendations from corresponding authors [supplementary materials A].

Table 2. ***Review Questions and Inclusion and Exclusion Criteria***

	Review questions	
	Do different sub-groups of adults accessing psychotherapy respond to treatment at different rates in line with the "Good enough level" perspective? Is the shape of change linear or non-linear?	
	Inclusion criteria	**Exclusion criteria**
Population	People over 16 accessing psychotherapy treatment.	Studies researching children and/or adolescents under 16.
Intervention	Any form of psychological intervention, delivered in any format.	Studies that do not include psychological interventions.
Comparator	Study design must stratify cases by treatment length and examine associations between treatment duration and outcomes based on the GEL concept directly.	Studies where cases are not compared by treatment length, for example only examining aggregate group responses to identify rates of change.
Outcomes	Response to psychotherapy 'dose' measured using standardized outcome measures, examining the rates of change.	Studies that do not use standardized outcome measures or measure outcomes as a result of non-psychological interventions. Studies that do not examine either rate or shape of change in response to psychotherapy.
Setting	Any settings where psychological interventions are usually delivered, across clinical and non-clinical settings (including outpatient, inpatient, university counseling centers, etc.), in any country.	Non-psychological intervention settings.
Study design	Practice-based naturalistic studies or controlled trials of psychological interventions. Cases must be stratified by treatment length. Studies published in English in peer reviewed scientific journals.	Studies that do not use a stratified design (by treatment length). Literature not published in peer reviewed scientific journals. Research studies not in published in English.

Data Extraction

A standardized data extraction form gathered information on study aims, setting, sample size, demographics, inclusion/exclusion criteria, presenting problem, intervention, outcome measures, outcome criteria, methods, treatment duration, and key findings.

Risk of Bias Assessment

Risk of bias was assessed using the Critical Appraisal Skills Programme Cohort Study Checklist (CASP, 2018). Two further questions were added based on Cochrane library guidance relating to selective reporting and missing data (Higgins & Green, 2011). Ratings of eligible studies were completed independently by two reviewers (the lead author and a trainee clinical psychologist), and Cohen's Kappa was used to assess inter-rater reliability (Altman, 1999). [Supplementary materials B]

Data Analysis

The included studies examined the GEL in four different ways: (a) associations between improvement and treatment duration, (b) associations between baseline symptom severity and treatment duration, (c) assessing rates of change, and (d) assessing the shape of change [Supplementary materials C]. A narrative synthesis of findings is presented, organized according to these different methodological approaches. Random effects meta-analyses were also performed where sufficient data were available, using the statistical package Meta-Analysis via Shiny (Hamilton, 2017). Heterogeneity was examined using the Q and I^2 statistics (Higgins & Thompson, 2002). Potential publication bias was examined using the weight-function likelihood ratio test (Vevea & Hedges, 1995) and the regression test for funnel plot asymmetry (Egger, Smith, Schneider, & Minder, 1997). There is debate as to whether small study numbers should be used in meta-analyses. Following the argument by Borenstein et al. (2009), we performed meta-analysis so as to enable evidence-based conclusions guided by any available data, taking care to identify and report indices of heterogeneity that may influence the interpretation of results. We pre-registered our plan to carry out random effects meta-analysis on this basis.

Results

Study Characteristics

Figure 1 summarizes the search and study selection process. A total of *K*=2, 299 records were initially identified. One additional eligible study was obtained via correspondence with authors of selected studies, and *k*=2, 083 were left after removing duplicates. Following screening of titles abstracts and full-texts, *k*= 15 papers were included in the review.

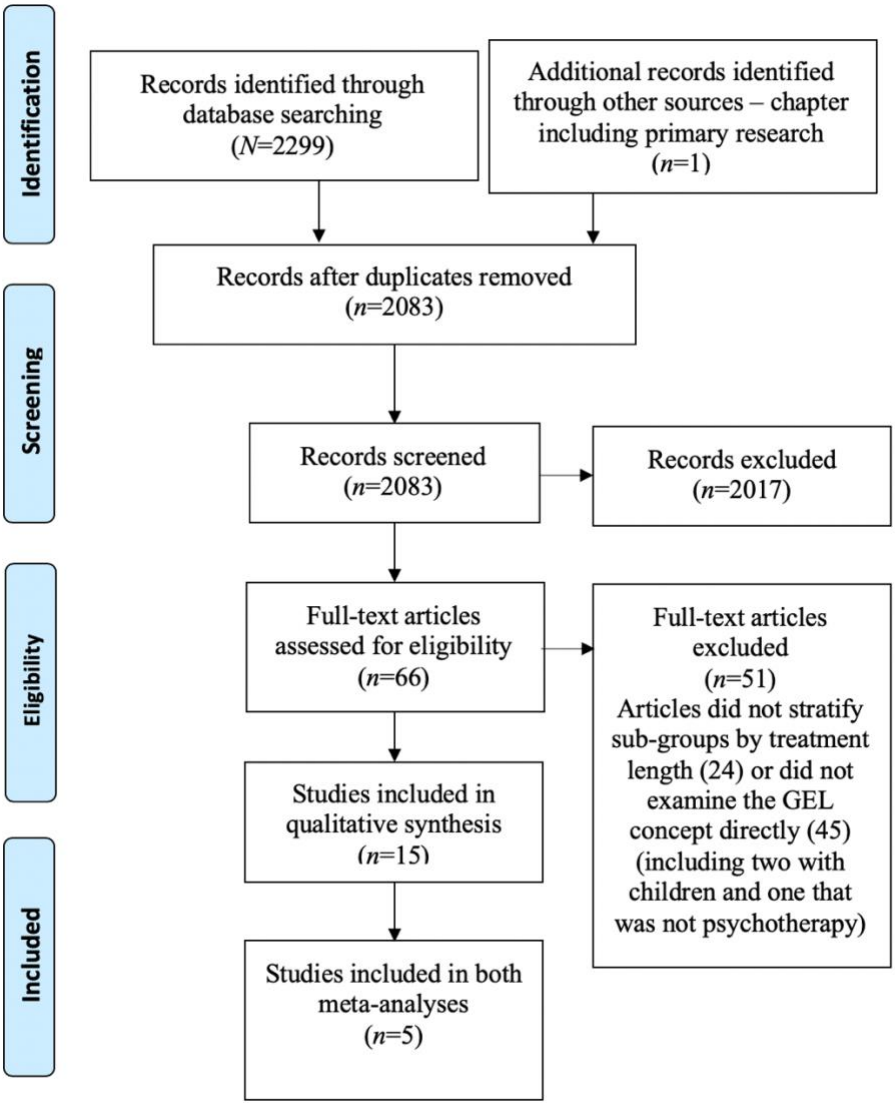

Figure 1. Prisma diagram based on Moher, Liberati, Tetzlaff, and Altman, 2009.

Design, setting and sample size. Table 3 summarizes the characteristics of eligible studies, most of which (*k*= 14) were analyses of naturalistic psychotherapy outcomes data, and one applied random allocation of clients to fixed treatment lengths (Barkham et al., 1996). Five studies were UK-based (mixed settings), nine were US-based (all university counseling centers apart from one community center), and one from Sweden (primary and psychiatric samples). The total sample across studies was *N*= 204,901, with *n*=114,123 included in the main GEL analyses.

Table 3. ***Study Characteristics***

First Author and Year	Study Design	Study Setting	Presenting Problems	Total *N* (204,901)	Analyzed *n* (114,123)	Intervention	Outcome Measures/criteria	Duration
1. Baldwin et al. (2009)	Database analysis	US University counseling center	Mixed	4676	2985 above cut-off	Mixed	OQ-45 RCSI	Mean 6.46 sessions
2. Barkham et al. (1996)	Random allocation	UK Psychotherapy settings	Mixed, with 85% depression	212	106 in 8 105 in 16	CBT or PI	BDI, IPP-32, PQ	Fixed, 8 or 16 sessions
3. Barkham et al. (2006)	Database analysis	33 UK NHS Primary care	Mixed	1868	1472 above cut-off	Mixed	CORE-OM RCSI/RC	Some fixed but flexible, PE, 12 sessions or less
4. Erekson et al. (2015)	Database analysis	US University counseling	Mixed	22,235	21488	Mixed	OQ-45 RCSI	Mean 5.8 sessions
5. Evans et al. (2017)	Database analysis	UK Secondary care	Mixed	4877	925	Mixed	CORE-OM RC	Median 15 sessions, 26 weeks, .61 per week
6. Falkenström et al. (2016)	Database analysis	Swedish Primary and psychiatric services	Mixed	1794	924	Mixed	CORE-OM Scores modelled	Mean 6 primary care / 9.1 psychiatric
7. Gottfredson et al. (2014)	Database re-analysis (Baldwin et al. 2009)	US University counseling	Mixed	4676	2985	Unknown	OQ-45 Scores modelled	Median 8 sessions/6.89 weeks
8. Kivlighan et al. (2019)	Database analysis	US University counseling	Unknown	786	438 / 369 with ending info	Unknown	BHM-20 Scores modelled	Some PE. Mean 5.54 sessions

9.	Nielsen et al. (2016)	Database analysis	US University counseling	Mixed	24,860	17,490	77.8% individual, then mixed.	OQ-45 RC	Median 4, modal 1 (1-548)
10.	Owen et al. (2015)	Database analysis	47 US College counseling centers & 1 community center	Unknown	38,985	10,854	Unknown	BHM Scores modelled	Mean 9.41, median 8 sessions
11.	Owen et al. (2016)	Database analysis	46 US College counseling centers & 1 community center	Unknown	48,963	13,664	Unknown	BHM RC / scores modelled	Mean 9.04 sessions
12.	Reese et al. (2011)	Database analysis	US University counseling	Mixed	3270	1207	Mixed	OQ-45 Scores modelled	90% <15 sessions, median 5
13.	Stiles et al. (2008)	Database analysis	UK 32 Primary care services	Mixed	9703	9703	Mixed	CORE-OM RCSI / mean change	PE, <=20 sessions. Some fixed=6 but flexible
14.	Stiles et al. (2015)	Database analysis	UK NHS 6 Primary care, 8 secondary care, 2 tertiary care, 10 University, 14 voluntary, 2 private	Mixed	36,297	26,430	Mixed	CORE-OM RCSI	PE, Some fixed (6) but flexible, median 6 sessions.
15.	Stulz et al. (2013)	Database analysis	US 20 College counseling centers, 4 primary care centers, 2 private centers.	Mixed	6375	6331	Mixed	BHM RCSI	Median 5 sessions

Overlapping samples. There was some reported overlap in the samples. Gottfredson, Bauer, Baldwin, and Okiishi (2014) provided a re-analysis of data from Baldwin, Berkeljon, Atkins, Olsen, and Nielsen (2009), however this examined the impact of missing data and is not aggregated in results sections. Stiles, Barkham, and Wheeler (2015) reported that there may be up to 1.8% data overlap between their study and Stiles, Barkham, Connell, and Mellor-Clark (2008), and Barkham et al. (2006). The data from these studies was aggregated in meta-analyses however the impact of this overlap is considered to be low. There was also database overlap between Owen et al. (2015) and Owen, Adelson, Budge, Kopta, and Reese (2016). However, the latter studies examined different aspects of the GEL model and are not treated as unique samples for aggregation here.

Measures. Six outcome instruments were used across studies, including measures of depression (Beck Depression Inventory [BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961]), interpersonal functioning (Inventory of Interpersonal Problems [IIP-32; Barkham, Hardy, & Startup, 1996]), ideographically defined problems (Simplified version of the Personal Questionnaire [PQ; See Mulhall, 1976 - originally developed by Shapiro, 1961]), and measures of general psychological distress and functioning (Clinical Outcomes in Routine Evaluation – Outcome Measure [CORE-OM; Evans et al., 2002]; Outcome Questionnaire-45 [OQ-45; Lambert et al., 1996]; Behavioral Health Measure [BHM; Kopta & Lowry, 2002]).

Outcome criteria. All studies used either the concept of reliable change (RC) or that of reliable and clinically significant improvement (RCSI). RC refers to a client's pre-post treatment change that has not occurred by chance, and is calculated using the standard error of difference for a particular measure (Jacobson & Truax, 1991). RCSI refers to both achieving criteria for RC and seeing scores that move from clinical to non-clinical thresholds,

as defined by population norms for those particular measures (Evans, Margison, & Barkham, 1998).

Interventions. A wide variety of interventions were reported, including cognitive behavioral therapy, psychodynamic interventions, and integrative approaches. Most studies had limited information about the psychological therapies employed.

Risk of Bias Assessment

All of the studies were considered to have relatively low risk of bias. Cohen's Kappa found moderate agreement between raters, $k = .51$, $p < .001$ (Altman, 1999), where ratings matched 85% of the time. In discussion, the majority of disagreements were on whether authors had identified and overcome all confounds ("yes" versus "unclear") and whether there were unaccounted for missing data. Disagreements were discussed and resolved without the need for mediation by a third reviewer.

Narrative Synthesis

Four approaches to examining the GEL model were identified in the literature: (a) associations between improvement and treatment duration, (b) associations between baseline symptom severity and treatment duration, (c) assessing rates of change, and (d) assessing the shape of change. Key findings from all reviewed studies are documented in Table 4 and methods are described in supplementary materials C and D.

Table 4. ***Findings Reported by Approach and Method Used***

First Author and Year	Method	Reported Findings/Statistics
	Associations between improvement and total sessions	
Baldwin et al. (2009)	Logistic regression using total sessions as predictor of RCSI. Min=3 sessions. RCSI binary. Correlation between sessions totals and final scores.	Small non-linear relationship between RCSI and total sessions – small increase up to session 8, then rates of RCSI plateau. Loglinear term significant for sessions and RCSI, odds ratio: 3.08, p<.05. Converted to *r*=0.2962 for meta-analysis. No correlation between sessions and final scores *r*=.02, *p*=.09.
Barkham et al. (2006)	Percentage calculation of RCSI per group. Correlation between rate of RCSI and total sessions.	Large negative correlation between rates of RCSI and total sessions *r*=-.91, p<.001 (up to 12 sessions).
Evans et al. (2017)	Correlation between change in score and total sessions. Min=3 sessions. Examined differences between reliable change categories and dose.	No correlation between change in score and total sessions r_s= -.04, p=.289. No significant differences between reliable change groups and total sessions, *H*(3)=.67, *p*=.879.
Owen et al. (2016)	Regression between amount of change on items and total sessions.	Small associations on individual items: Wellbeing: r^2=.014; Symptom distress: r^2=.021; Life functioning: r^2=.004.
Nielsen et al. (2016)	Linear correlation between change and total sessions. Linear and non-linear regressions using various terms between change scores and total sessions. SEMs to analyze regressions of symptom change on sessions (sessions predict change - DR) and sessions on change (change predicts sessions - GEL). Plus a combined DR and GEL SEM. Analyzed with X^2.	No linear correlation *r*=.008, p=.29. However inverse (NAC) regression significant: *F*(1, 17488)=72.5, *p*<.001, R^2 =.004. Increases in change seen up to session 18 then plateaus. When reliable change criteria is used, plateau occurs at 6 sessions. SEMs showed that the only adequate fit was achieved by a DR *plus* GEL SEM: X^2(1, *n*=17490)=2.5, *p*=.065. Variance explained was improved by individual therapy modality effects (.02% to 13%).
Stiles et al. (2008)	Percentage calculation of RCSI per group Correlation between RCSI / RC and total sessions. Compare mean pre-post change scores by total sessions.	Change scores similar across treatment lengths. Large negative correlation between RCSI and total sessions. No correlation between RC and total sessions. RCSI: *r*= -.75, p<.001. RC: *r*=.11, ns.
Stiles et al. (2015)	Percentage calculation of RCSI per group Correlation between rates of RCSI / RC and total sessions. Compare mean pre-post change scores by total sessions.	Change scores similar across treatment lengths. Large negative correlation between RCSI and total sessions. Moderate negative correlation between RC and total sessions. RCSI: *r*=-.58, p<.001. RC: *r*=-.40, p<.001.
Stulz et al. (2013)	Correlation between rates of RCSI and total sessions. Min=3 sessions.	Large positive correlation between RCSI and total sessions *r*=.714, p=.004.

Associations between baseline symptom scores and total sessions		
Baldwin et al. (2009)	Correlation between baseline score and total sessions. Min=3 sessions.	Small positive correlation between baseline score and total sessions. *r*= .09, p<.001.
Barkham et al. (2006)	Correlation between baseline score and total sessions.	Small positive correlation between baseline score and total sessions. *r*= .13, p<.001.
Erekson et al. (2015)	MLM with linear, quadratic and cubic terms. Min=2 sessions.	Higher levels of dose associated with lower levels of OQ-45 at intercept.
Evans et al. (2017)	Correlation between baseline score and total sessions. Min=3 sessions.	Small-moderate positive correlation between baseline score and total sessions. *r*= .29, p<.005.
Falkenström et al. (2016)	MLGMs comparing DR and GEL models to assess whether rate of change varies as function of treatment length. Min=3 sessions.	Although they found that initial symptom severity was not related to treatment length in weeks, the psychiatric sample had higher risk and higher total sessions numbers.
Owen et al. (2015)	3-level model, initial scores nested in clients nested in therapists. Min=4 sessions.	Clients in different classes showed differences in intake scores – 'Early & Late', and 'Slow & Steady', had higher intake scores than 'Worse Before Better'. Slow & Steady more distressed and slower trajectory overall.
Stiles et al. (2008)	Correlation between baseline score and total sessions. Correlation between mean baseline scores and total sessions.	Small positive correlation between baseline score and total sessions. *r*=.16, p<.00. Large positive correlation between mean baseline score and total sessions *r*=.93, *p*<.001.
Stiles et al. (2015)	Correlation between baseline score and total sessions. Correlation between mean baseline scores and total sessions.	Small positive correlation between baseline score and total sessions. *r*=.08, p<.001. Large positive correlation between mean baseline score and total sessions *r*=.58, *p*<.001.
Assessing rates of change		
Baldwin et al. (2009)	MGCM – compared average rate of change with total sessions. Min=3 sessions.	Significant interaction between rate and dose, slower rates associated with higher dose. Log of total sessions and cubic form: cubic (beta): 0.02, *p*<.01. Interactions between log of total sessions and time: Linear =2.69, Quad= -.29, Cubic=.02, all p<.01.

Barkham et al. (1996)	Percentage calculation of RCSI per group (8 or 16 sessions).	8 session group had faster rates of improvement than 16 session group at 8 sessions on BDI (X_2(1, n=181)=6.03, p=.014) and PQ items. However not on IPP-32. On BDI – faster reductions in distress, slower in characterological/interpersonal. Explains slower rates on IPP, also seen in PQ items.
Erekson et al. (2015)	MLM with total sessions and session frequency as continuous variable on rate of change. Min=2 sessions.	Higher doses had slower improvement rates, less frequent sessions had slower rates of change. Adding session frequency improved BIC by 8,515. Rate of change (based on clinically significant change) was faster in weekly than fortnightly groups based on total sessions: X_2= 39.36(1), p<.001. Effect size of session frequency f_2 0.07.
Falkenström et al. (2016)	MLGMs comparing DR and GEL models to assess whether rate of change varies as function of treatment length. Min=3 sessions.	GEL model a better fit in primary (X_2(4) = 37.46, p<.001) and psychiatric (X_2(3) =25.68, p<.001) samples. Faster rates of change with fewer sessions in both samples, but psychiatric saw slower rates of change and higher total sessions.
Gottfredson et al. (2014)	SPMMs used to re-analyze data from Baldwin et al. (2009), to handle missing data.	SPMMs indicated that faster responders were more likely to terminate therapy earlier, meaning rates of change underestimated (6.50% - 6.66% across two models).
Kivlighan et al. (2019)	MLM estimated with linear, log-linear and quadratic terms – measure broken down into different domains and dependency between items controlled for. Min=2 sessions. Analyzed planned vs unspecified endings.	Log-linear best fit for all >=2 sessions, linear best fit for all >=3 sessions. Rate of change did not vary on individual domains, but did overall: (-0.01, p = .024). People more likely to terminate early due to changes in wellbeing but not other items.
Owen et al. (2015)	GMM. Identified 3 different classes (1. Early and late, 2. Worse before better, 3. Slow and steady). Modelled linear, quadratic and cubic rates of change. Min=4 sessions.	All were significant, initial rates of change (over first 3 sessions) differed – slow and steady class had slower rate of change than early and late, and worse before better. Coefficients on initial rates of change: Slope Class 3 vs Class 2: 22.75, Class 1 vs Class 3: 4.93, p<.001.
Owen et al. (2016)	MLMs estimated rate of change for DR and GEL models and compared fit. Min=1 session. On individual questionnaire domains.	GEL Log-linear model was best fit for wellbeing and symptom distress (Loglinear x sessions interaction coefficients: -0.0098 / -0.0081, p<.01). GEL quadratic model best fit for life functioning ($Session_2$ x sessions interaction coefficient: 0.0002, p<.01). Clients attending fewer sessions had faster rates of change. However change on life functioning was smaller than wellbeing or symptom distress. Therapist effects explained some of variations in change on wellbeing and life functioning.

Reese et al. (2011)	MLGM with improvement as a function of total sessions and session frequency. Used linear, cubic and quadratic terms.	GEL model significantly better fit than DR, longer sessions had slower rates of change. GEL modified (including session frequency) was significantly better fit than GEL alone, less frequent sessions had slower rates of change. GEL: X2(2)=98.2, p<.001. GEL vs GEL mod:X2=18.1, p<.001. Overall linear trends most parsimonious – linear and steeper at <5.72 sessions.
Stulz et al. (2013)	LGCMs – correlated mean rates of change with total sessions. Min=3 sessions.	Large negative correlation between mean change and total sessions: r= -.974 (for log-linear model – best fit).
	Assessing shape of change	
Baldwin et al. (2009)	MGCMs compared DR and GEL, modelled as linear based on previous studies then cubic based on visual inspection. Measures every session. Min=3 sessions.	DR model produced NAC, however GEL model fit with cubic terms superior (double curve) $X_2(4)$=428.49, p<.0, Cubic beta= -.06, p<.01. Cubic BIC: 244,425
Barkham et al. (1996)	Percentage calculation of RCSI per group Pre, mid (for 16 sessions), and post therapy.	Linear improvement seen on PQ items and in sequence of RCSI percentages on BDI or IPP. When aggregated across both groups however Log-linear NAC shape seen.
Erekson et al. (2015)	MLM with linear, quadratic and cubic terms. Min=2 sessions.	All significant but linear largest estimate.
Falkenström et al. (2016)	MGLMs comparing DR and GEL models using linear, quadratic and cubic terms. Min=3 sessions.	GEL model a better fit in primary ($X_2(4)$ = 37.46, p<.001) and psychiatric ($X_2(3)$ =25.68, p<.001) samples. In primary care: Linear, cubic and quadratic all significant but quadratic shape best. In psychiatric sample linear shape best.
Kivlighan et al. (2019)	MLMs estimated with linear, log-linear and quadratic terms – measure broken down into different domains and dependency between items controlled for. Min=2 sessions.	Log-linear best fit for all >=2 sessions (BIC 35,728.83), linear best fit for all >=3 sessions (BIC 3320.65).
Nielsen et al. (2016)	Linear and non-linear terms used in regression analyses of change scores and total sessions. Then used SEM to identify more complex relationships between shape of change and whether total sessions predict improvement or improvement predicts total sessions.	Inverse (NAC) regression significant/largest: F(1, 17488)=72.5, p<.001, R_2 =.004. Increases in change scores seen up to session 18 then plateaus. When criteria of reliable change is used, rates plateaued by the 6th session. Higher sessions fit GEL, shorter fit DR. Combined DR and GEL SEMs fit data best.

Owen et al. (2015)	GMM to identify sub-classes. Modelled linear, quadratic and cubic forms. Min=4 sessions.	3 classes model significant: Class 1 = early and late change (largest), Class 2=worse before better (smallest), Class 3=slow and steady (linear, longer therapy). AIC: 1, 087, 760. Adjusted BIC: 1, 087, 957.
Owen et al. (2016)	MLMs – Compared fit for log-linear, cubic and quadratic terms for DR and GEL models. On individual questionnaire domains. Measures every session. Min=1 sessions.	GEL better fit than DR. GEL Log-linear model was best fit for wellbeing and symptom distress, quadratic on life functioning. Clients having fewer sessions saw log-linear trend, those having longer sessions saw more linear trend. Wellbeing: GEL Log-linear BIC: 201, 622. Symptom distress: GEL Log-linear BIC:121,483. Functioning: GEL quadratic BIC: 174,939.
Reese et al. (2011)	MLGMs - compared aggregate, GEL, and GEL with session frequency. Used linear, cubic and quadratic terms. Measures every third session. Min=1 session.	GEL with session frequency best fit. The GEL model also explained 3% more variance in scores than DR. Cubic terms significant but non-linear trend very subtle so linear terms used. GEL vs GEL modified: $X_2(2)=18.1$, $p<.001$ GEL modified AIC=30, 709.4. Overall linear trends most parsimonious – linear and steeper at <5.72 sessions.
Stulz et al. (2013)	LGCMs – compared linear and log-linear stratified models. Min=3 sessions. Measures every session	Log-linear outperformed linear regardless of treatment length. (Online supplement figures not available).

Notes. Where studies refer to comparisons between the DR model and the GEL model, they mean aggregated or stratified by total sessions received.
Min.=3 for e.g., refers to minimum number of sessions. Model abbreviations: MGCM: Multi-level growth curve model. MLM: Multi-level model. LGCM: Latent growth curve model. MLGM: Multi-level growth model. GMM: Growth mixture model. SEM: Structural equation modeling. SPMM: Shared parameter mixture model.

(a) Associations between improvement and treatment duration. Eight studies examined this relationship, with six using correlation analyses and two using regression. Five studies found support for the GEL model, reporting no –or negative– correlations between improvement and total sessions (Studies: 3, 5, 9, 13, 14). Two of these studies also compared mean change scores by total sessions, finding similar change scores regardless of treatment duration (13, 14). Two studies found small associations (1, 11) and Stulz et al. (2013) found a large positive correlation. Using structural equation modeling to investigate the direction of associations between treatment duration and outcomes, Nielsen et al. (2016) reported that the best fit for their data was attained using a combined DR and GEL model. Treatment duration could predict change, but only in a model where it was also possible for change to predict duration.

(b) Associations between baseline symptom severity and treatment duration. Eight studies examined associations between initial symptom severity and treatment duration. Five of these reported significant positive correlations, suggesting that people with higher baseline severity tend to have longer treatments (1, 3, 5, 13, 14). One further study (6) applied multilevel growth linear models to compare primary care and psychiatric samples, finding that the psychiatric sample had higher severity and higher treatment duration. One study (10) used growth mixture modeling to show that higher baselines were associated with different sub-classes of clients, in particular those showing “early and late changes”, or “slow and steady” progress. One study (4) however reported that higher levels of dose were associated with lower OQ-45 scores at intercept.

(c) Assessing rates of change. Nine studies assessed whether rates of change differ depending on treatment length (1, 2, 4, 6, 8, 10, 11, 12, 15). All nine studies reported that rates of change on global scores were faster in cases that had fewer sessions. Two studies (4,

12) expanded on this by showing that those having more frequent sessions had faster rates of change (e.g., more than one per week). Furthermore, two studies (2, 11) found that problems relating to characterological, interpersonal or life functioning factors appeared to respond slower than problems relating to wellbeing or symptom distress.

Although Kivlighan, Lin, Egan, Pickett, and Goldberg (2019) found that rates of change varied as a function of total sessions on global distress scores on the BHM-20, they found no difference in rates of change when sub-domains were examined and item dependency was controlled for. They further report that early termination from treatment was associated with improvements on wellbeing but not on other domains (symptom distress or life functioning).

Owen et al. (2016) describe that therapist effects explained some of the variance in rates of change in wellbeing and life functioning in their study, and Owen et al. (2015) noted that different sub-classes of clients responded at different rates; notably the "slow and steady" group had the slowest trajectories. Gottfredson et al. (2014) also reanalyzed data from Baldwin et al. (2009) using shared parameter mixture models to handle "non-ignorable" missing data, suggesting that rates of change may also be underestimated using typical methods.

(d) Assessing the shape of change. Ten studies examined the shape of change. One study described the shape of change based on visual inspection of plots of scores (2), and nine assessed the model fit of linear, log-linear, quadratic or cubic shapes of change (1, 4, 6, 8, 9, 10, 11, 12, 15). Eight of these found a variety of shapes of change and reported on the overall best fit for their data. Five studies contrasted a DR model (aggregating samples) with a GEL model (stratifying samples) (1, 6, 9, 11, 12) and all of them found the GEL model to provide better goodness-of-fit; as described, Nielsen et al. (2016) noted that a combined model had even better fit.

Linear trends. A linear shape of change was the best fit in seven studies under certain conditions. Barkham et al. (1996) described that change looked linear when broken down into different symptoms, on individualized items, or when comparing sequences of RCSI rates. Reese, Toland, and Hopkins (2011) used multilevel growth linear models and found that although a cubic term was significant, linear trends described the data more parsimoniously. Similarly, Erekson et al. (2015) found a linear shape most representative of their sample. Four studies comparing sub-groups found linear terms to offer the best fit at longer treatment lengths. Kivlighan et al. (2019) describe a linear pattern in clients having three or more sessions, as opposed to log-linear patterns evidenced in those having two or more. Falkenström et al. (2016) found a linear shape in a psychiatric sample with longer treatments and slower rates of change, when compared with a quadratic trend seen in a primary care sample. Owen et al. (2016) described linear trends in those having longer treatments, whilst Owen et al. (2015) observed linear trends in a "slow and steady" sub-group who had longer treatments (note possible sample overlap in the latter two studies).

Log-linear trends. Four studies found log-linear trends in certain circumstances. For example, Stulz, Lutz, Kopta, Minami, and Saunders (2013) stratified groups by treatment length, finding that log-linear terms fit better than linear in their sample, regardless of treatment length. Kivlighan et al. (2019) examined shapes of change for those having greater than two sessions versus those having greater than three, finding a log-linear shape in those with at least two sessions compared with a linear shape in those with at least three. Owen et al. (2016) found that a log-linear trend offered the best fit for the problem domains of wellbeing and symptom distress but not life functioning (which was quadratic), as well as for those having shorter treatments. Nielsen et al. (2016) observed a log-linear trend in their data according to visual inspection and regression terms. They described that a log-linear trend fit better for shorter treatment lengths, whilst a linear model fit better in longer treatments

lengths. Using structural equation modeling, they found that a combined DR and GEL model offered the best overall fit.

Quadratic trends. Two studies found quadratic trends in certain circumstances. Owen et al. (2016) found this trend on the problem domain of life functioning. Falkenström et al. (2016) found that a quadratic trend best described a primary care sample, whilst a linear term better described the psychiatric sample.

Cubic trends. A cubic trend was found to offer the best fit in two studies (Baldwin et al., 2009, and Reese et al., 2011). However, Reese et al. (2011) stated that on visual inspection the trend was better described as linear. Owen et al. (2015) also found an "early-and-late" change trend in their largest sub-class of clients, resembling a cubic trend.

Meta-Analysis

Five studies reported correlation coefficients for associations between baseline severity and treatment duration (measured in sessions), and five reported correlation coefficients for associations between treatment duration and outcomes (reliable and clinically significant improvement [RCSI]). Two meta-analyses were therefore carried out to examine pooled correlation coefficients using a random effects model (see supplementary materials).

Associations between initial symptom severity and treatment duration. Five studies (1, 3, 5, 13, 14; n= 41,515) were included all of which reported positive correlations between baseline symptom scores and total sessions attended (ranging from 0.08 to 0.28 – see supplementary materials E). A significant small pooled effect size of r = 0.15 [95% CI = 0.08, 0.22], $p < .001$ was found, suggesting that higher baseline severity was associated with longer treatment. However, high heterogeneity was indicated ($Q(4) = 83.20$, $p < .001$), with I^2 of 95.2%. Publication bias analysis was non-significant according to the weight-function $x^2(1) = 1.08$, $p = 0.29$, and funnel plot tests $t = 1.41$, $p = 0.25$.

Note that the study showing the highest correlation (Evans et al., 2017) used data from UK secondary care services as opposed to primary or university counseling services. There were also three other studies examining symptom severity and duration, which were not possible to combine for quantitative analysis: one found a negative association, one found a positive association only in a psychiatric sample, and one found positive associations in particular sub-classes. Also note that when mean rather than individual baseline scores were used in Stiles et al. (2008), a larger positive correlation was found. This may be explained by the heterogeneity of individual baseline scores.

Associations between treatment duration and clinical outcomes. Five studies (n = 46,921) were included (1, 3, 13, 14, 15). Using the criteria of RCSI, a non-significant pooled effect size of r = -0.24 [95% CI = -0.70, 0.36], p = 0.27 was found, suggesting no linear correlation between treatment duration and outcome. However, this analysis combined results derived from three studies showing large negative correlations and two studies showing small-to-moderate or large positive correlations. As a consequence, high heterogeneity was indicated ($Q(4)$ = 18,655.94, p < .001), with I^2 of 100%. Publication bias analysis was non-significant according to the weight-function model likelihood ratio test $x^2(1)$ = 0.23, p = 0.64, and the regression test for funnel plot asymmetry $t(3)$ = 1.04, p = 0.37.

Sources of heterogeneity. Studies were examined for differences in criteria reported and potential sources of heterogeneity. Although high heterogeneity is to be expected across studies of varying treatment duration etc., a clear pattern was also observed relating to whether studies included planned or unspecified endings. Of the five studies examining RCSI and treatment duration, the three that included planned endings only (completers analysis) produced large negative correlations (3, 13, 14) whereas the two including unspecified endings (intention-to-treat analysis) found small-to-moderate (1) and large positive (15)

correlations. Further sub-group analyses were therefore performed depending on whether the studies included planned endings exclusively or whether ending information was unspecified.

Completers sub-group analysis. Three studies (3, 13, 14) were included with n = 37,605 participants. All three noted that some of the services included tended to limit therapy to six sessions (but not all), with flexibility to add more. A significant large pooled effect size of r = -0.63 [95% CI = -0.73, -0.51], p <. 001 was found, suggesting a negative correlation between recovery and total sessions when planned endings only are included. However high heterogeneity was again indicated $Q(2) = 1546.61$, $p < .001$), with I^2 of 99.9%. Although these studies all suggested a negative correlation between RCSI and total sessions, there were significant discrepancies between their effect sizes. Publication bias analysis was nonsignificant according to the weight-function model $X^2(1) = 4.571$, $p = 1$ and funnel plot test $t(1) = -2.387$, p = 0.253.

Intention-to-treat sub-group analysis. Two studies (1, 15) were included n = 9316. A significant moderate-large pooled effect size of r = 0.47 [95% CI = 0.10, 0.72], p = 0.042 was found. However high heterogeneity was indicated $Q(1) = 705.95$ $p < .001$), with I^2 of 99.9%. Publication bias analysis was non-significant, with a weight-function test of $X^2(1) = 0.05$, p = 0.824.

Note that the two studies finding positive correlations used data from US counseling services. The three studies finding negative correlations originated in the UK and had up to 1.8% overlap. Two were based in primary care (Barkham et al., 2006; Stiles et al., 2008) and one in mixed settings (Stiles et al., 2015). The mixed settings study found the smallest negative correlation between RCSI and total sessions (-0.52). One further UK study (Evans et al., 2017) examined change scores (rather than RCSI) using secondary care data, finding no association between total sessions and change in scores in this context.

It is possible that larger effects are produced dependent on the criteria used (e.g. RCSI produces stronger effects than RC due to the stricter criteria used, where slow or non-responders may be less likely to see RCSI than RC). It may also depend on the sample selected (e.g. based on complexity). However further research is needed to examine this as there were also differences in positive correlations between US counseling services without clear cause.

Discussion

Main Findings

This is the first comprehensive synthesis of the GEL literature, using systematic review and meta-analysis methodology. We found partial support for key assumptions of the GEL model. For example, baseline severity was significantly associated with therapy duration. This supports the notion that some people may require lengthier interventions than others, depending on symptom severity. Studies included in the meta-analyses were highly heterogeneous in accordance with a key assumption of the GEL model, which is that therapy duration is highly variable across samples. This was further supported by the highly heterogeneous findings across studies that examined rates and shapes of change, where linear change trends were supported in some samples and nonlinear trends in others. Put simply, the reviewed evidence indicates that different people change at different rates, and in some instances, this is associated with baseline symptom severity.

Although severity was significantly correlated with therapy duration, the present meta-analysis indicates that this association is weak (r = 0.15). However, this may be influenced by study setting, where secondary care, psychiatric samples and sub-group analyses were indicative of positive associations. It is also theoretically plausible that initial severity is a fairly crude proxy indicator of “complexity”, a concept that has been proposed to be influenced by multiple variables (symptom severity, personality, socioeconomic and

cultural features, etc.) that are statistically associated with treatment response (Delgadillo, Huey, Bennett, & McMillan, 2017). Our interpretation of the reviewed data is that less complex cases tend to have rapid response to treatment, whereas more complex cases with features associated with poorer outcomes, may require lengthier or more responsive interventions. As such, baseline severity indexes only one facet of the wider concept of "complexity", and weak statistical associations with treatment duration are unsurprising.

Evidence regarding the association between treatment duration and outcomes was mixed. Overall there was some support for the GEL model: pooling data across reviewed studies suggested no significant relationship between treatment duration and outcomes, and most studies found that rates of change varied as a function of total sessions. However, we cautiously draw attention to the relevance of study design. Different findings were observed depending on whether studies included or excluded cases that dropped out of treatment. Studies analyzing data for treatment completers tended to observe no -or negative- correlations between duration and outcomes, whereas studies including data for drop-out cases tended to find positive correlations.

Our reading of this is that when unplanned endings are included, samples are more likely to include those who drop out early before criteria for improvement have been met (thus suggesting an increased effect of therapy with dose). When studies include only treatment completers it is likely that therapy has continued until a good-enough level has been reached at a variety of durations (so the effect of therapy may look equivalent at a variety of treatment lengths). In this way, the two models capture a different focus: the GEL model better captures the heterogeneity of individual responses to therapy (for those who remain in therapy), whereas the DR model reflects a broader overall picture of responses to therapy across patients who complete and those who drop out of treatment. This may also be influenced by country of origin (and service models used), change criteria and complexity of

cases, although further research is needed to understand the influence of these sources of heterogeneity.

There was some support for the curvilinear relationship described by the DR model. This was most often found in those having shorter treatment lengths, whilst linear shapes were more likely to be found at longer treatment lengths. However, there were also differences in how this was examined with some studies aggregating findings into low and high treatment groups rather than stratifying by treatment length. It was clear that although different people responded more or less rapidly, most treatment responders tended to be identified within a time-limited boundary in these contexts (usually under 20 sessions) and the mean number of treatment sessions tended to be fairly low (see Table 3). This is partly consistent with the DR model concept of an *optimal dose*: even if the dose of treatment does not *cause* improvement, most cases that improve can be identified within a predictable number of therapy sessions. Thus, from the perspective of individual patients we observe that the marked heterogeneity in the time taken to attain symptomatic improvements is associated in variable treatment durations (*responsive regulation*), but from a clinical population perspective it is clear that treatment response generally occurs within a predictable window of time (dose-response parameters or *boundaries*). Such a pattern of evidence could be described using the expression "*boundaried responsive regulation*", which captures elements from both the GEL and the DR models, recognizing that both perspectives hold some wisdom about patterns of change in psychotherapy.

Limitations

Most of the reviewed studies were subject to limitations that are common in naturalistic study samples, including issues related to missing data and unclear descriptions of samples and psychological interventions. Although missing data are often treated as *missing at random* in statistical analyses, this assumption may not be appropriate. For example,

Erekson et al. (2015) found that missing session data in their study were correlated with session frequency, total sessions and baseline symptom severity. Evans et al. (2017) showed that those with completed measures were more likely to be older, White British, and with lower baselines than those without. Gottfredson et al. (2014) illustrated that when imputation methods were used to handle "non-ignorable" missing data, participants with faster recovery rates terminated therapy earlier, meaning that rates of change are generally underestimated according to traditional "missing at random" assumptions. Of further note is the finding by Kivlighan et al. (2019) that rates of change did not vary on sub-scores as a function of total sessions when item dependency was controlled for on the BHM-20. Further research should therefore include assessments of the impact of "non-ignorable" missing data and control for sub-scale item dependency.

Most reviewed studies were retrospective analyses of practice-based data, and –as such– were reliant on the recording of demographic and treatment information by the included clinics. Although missing participant characteristics do not preclude the examination of treatment outcomes, they may limit interpretations of findings. For example, it would be of particular interest to characterize the features of clients who show rapid versus gradual or non-responses to therapy, and such analyses are dependent on the availability of client and therapist-level variables. Given that these studies reported different findings based on whether planned or unplanned endings were included, better recording of the reasons for treatment ending would also facilitate clearer interpretations of the GEL.

Issues related to missing data and scarce availability of information about clients, therapists and treatments may explain the high heterogeneity found across studies. We also note that a considerable proportion (but not all) of the GEL literature comes from studies including Caucasian student counseling or primary care samples, and their findings may not necessarily generalize to other clinical samples and settings. We cannot therefore assume that

the GEL model assumptions are broadly generalizable. In addition, although sample sizes across studies tended to be large, few studies provided sufficient statistical information for meta-analysis. Other limitations specific to the review methodology include the exclusion of studies written in languages other than English and the exclusion of grey literature. There may therefore be missed findings that could contribute to further analysis of the GEL. However, none of the current GEL authors and leaders in the field were aware of further missing literature that we could have included, and it was considered important that such technical literature had undergone expert peer review prior to inclusion.

Theoretical Implications and Future Research

Several key theoretical questions have emerged from this review. For example, if some people respond more rapidly to therapy than others, it is of interest to know if we can identify their profiles. Future research could help attain greater precision in the targeted allocation of brief versus lengthy psychological interventions, developing treatment selection algorithms using information from clients, therapists, and different outcome domains. It could therefore be possible to offer low intensity and low-cost therapies to those most likely to be rapid responders, and allocate gradual responders to more intensive treatment. Recent client-profiling studies have shown that this stratified allocation of low versus high intensity treatments has the potential to improve the effectiveness (Delgadillo et al., 2017) and efficiency of psychological care (Delgadillo et al., 2020).

As discussed, nine of the reviewed studies used data from university counseling centers in the US, and in the UK the majority of the research came from primary care clinics. It would therefore be of interest to understand if these findings generalize to other – potentially more complex– samples. Future studies could apply *multivariable prognostic indices* (e.g., see Delgadillo et al., 2017; Lorenzo-Luaces, DeRubeis, van Straten, & Tiemens,

2017) to investigate associations between case complexity and treatment duration, in a way that includes but moves beyond simple associations with baseline severity.

Some of the studies in this review also highlighted other influences on rates of change, such as session frequency and therapist effects (see also Goldberg, Hoyt, Nissen-Lie, Nielsen, & Wampold, 2018). Better reporting of client and therapist demographics, and clinic and therapeutic contexts, as well the inclusion of more diverse samples in research would facilitate not only an understanding of "who" is less likely to respond but also assist with interpretations of "why". It is also important to note that in practice the length of treatment may be highly influenced by the services system in the respective country rather than based on patient need (Flückiger, Wampold, Delgadillo, Rubel, Vîslă, & Lutz, 2020).

Finally, it would be of interest to gain insight into clients' views about the types of outcomes that might constitute a good-enough level of improvement. For example, Kivlighan et al. (2019) noted that some people made progress on aspects such as wellbeing and terminated treatment on that basis, before making progress on other symptoms. Research has begun to consider whether symptom reduction should always the goal of therapy, making the claim that better understanding of client-defined outcomes is necessary (Cuijpers, 2019). A question for future research therefore is: what constitutes a GEL, and how can this be captured meaningfully in research findings?

Conclusions

Overall, some evidence supported the GEL assumptions, but some assumptions from the DR model were also supported. To account for these mixed findings, we propose the notion of *boundaried responsive regulation*: individuals may show different patterns and rates of clinical improvement, yet this occurs within predictable boundaries consistent with the notion of an overall optimal dose of therapy. The implications of this are that clinics should be planned flexibly so that treatment can continue until a good-enough level of

improvement is attained, yet this is still proposed to be within the guidelines provided by the dose response literature.

References

*Indicates studies that were included in the systematic review

Altman, D. G. (1999). *Practical statistics for medical research*. New York, NY: Chapman & Hall/CRC Press.

*Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose-effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology*, *77*, 203–211. https://doi.org/10.1037/a0015235

*Barkham, M., Connell, J., Stiles, W., Miles, J., Margison, F., Evans, C., & Mellor-Clark, J. (2006). Dose-effect relations and responsive regulation of treatment duration: The good enough level. *Journal of Consulting and Clinical Psychology, 74*, 160-167. doi:10.1037/0022-006X.74.1.160

Barkham, M., Hardy, G.E., & Startup, M. (1996). The IIP-32: Development of a short version of the Inventory of Interpersonal Problems. *British Journal of Clinical Psychology*, *35*, 21-35. https://doi.org/10.1111/j.2044-8260.1996.tb01159.x

*Barkham, M., Rees, A., Stiles, W. B., Shapiro, D. A., Hardy, G. E., & Reynolds, S. (1996). Dose-effect relations in time-limited psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 64*, 927-935. http://dx.doi.org/10.1037/0022-006X.64.5.927

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561 - 571. doi: 10.1001/archpsyc.1961.01710120031004

Castonguay, L., Barkham, M., Lutz, W., & McAleavey, A. (2013). Practice-oriented research: Approaches and applications. In M. J. Lambert (Ed.). *Bergin and Garfield's*

handbook of psychotherapy and behavior change (6th ed., pp. 85-133). New Jersey, NJ: John Wiley & Sons, Inc.

Critical Appraisal Skills Programme (2018). *Cohort study checklist.* Retrieved from: https://casp-uk.net/casp-tools-checklists/

Cuijpers, P. (2019), Targets and outcomes of psychotherapies for mental disorders: an overview. *World Psychiatry, 18*, 276-285. doi:10.1002/wps.20661

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society, 11,* 121–136. https://doi.org/10.1080/15248371003699969 [supplementary materials]

Delgadillo, J., Appleby, S., Booth, S., Burnett, G., Carey, A., Edmeade, L., … & Lutz, W. (2020). The Leeds Risk Index: Field-test of a stratified psychological treatment selection algorithm. *Psychotherapy and Psychosomatics.* https://doi.org/10.1159/000505193

Delgadillo, J., Huey, D., Bennett, H., & McMillan, D. (2017). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology, 85,* 835-853. http://dx.doi.org/10.1037/ccp0000231

*Erekson, D. M., Lambert, M. J., & Eggett, D. L. (2015). The relationship between session frequency and psychotherapy outcome in a naturalistic setting. *Journal of Consulting and Clinical Psychology*, *83*, 1097–1107. https://doi.org/10.1037/a0039774

*Evans, L. J., Beck, A., & Burdett, M. (2017). The effect of length, duration, and intensity of psychological therapy on CORE global distress scores. *Psychology and Psychotherapy: Theory, Research and Practice, 90*, 389-400. https://doi.org/10.1111/papt.12120

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J. & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry 180,* 51-60. doi:10.1192/bjp.180.1.51

Evans, C., Margison, F., & Barkham, M. (1998). The contribution of reliable and clinically significant change methods to evidence-based mental health. *Evidence Based Mental Health, 1,* 70-72. http://dx.doi.org/10.1136/ebmh.1.3.70

*Falkenström, F., Josefsson, A., Berggren, T., & Holmqvist, R. (2016). How much therapy is enough? Comparing dose-effect and good-enough models in two different settings. *Psychotherapy*, *53*, 130–139. https://doi.org/10.1037/pst0000039

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). London, UK: SAGE Publications Ltd. [supplementary materials]

Flückiger, C., Wampold, B. E., Delgadillo, J., Rubel, J., Vîslă, A., & Lutz, W. (2020). Is there an evidence-based number of sessions in outpatient psychotherapy? A comparison of naturalistic conditions across countries. *Psychotherapy and Psychosomatics*, 1-3. https://doi.org/10.1159/000507793

Goldberg, S. B., Hoyt, W. T., Nissen-Lie, H. A., Nielsen, S. L., & Wampold, B. E. (2018). Unpacking the therapist effect: Impact of treatment length differs for high- and low-performing therapists. *Psychotherapy Research, 28,* 532-544. https://doi-org.sheffield.idm.oclc.org/10.1080/10503307.2016.1216625

*Gottfredson, N. C., Bauer, D. J., Baldwin, S. A., & Okiishi, J. C. (2014). Using a shared parameter mixture model to estimate change during treatment when termination is related to recovery speed. *Journal of Consulting and Clinical Psychology, 82*, 813-827. http://dx.doi.org/10.1037/a0034831

Hamilton, W. (2017). Package 'MAVIS' (Version 1.1.3). Retrieved from: http://kylehamilton.net/shiny/MAVIS/

Harnett, P., O'Donovan, A., & Lambert, M. J. (2010), The dose response relationship in psychotherapy: Implications for social policy. *Clinical Psychologist, 14*, 39-44. doi:10.1080/13284207.2010.500309

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (version 5.1.0). The Cochrane Collaboration. Retrieved from: www.handbook.cochrane.org

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539-1558. https://doi.org/10.1002/sim.1186

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinksy, D. E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist, 41*, 159-164. http://dx.doi.org/10.1037/0003-066X.41.2.159

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19. http://dx.doi.org/10.1037/0022-006X.59.1.12

Kadera, S. W., Lambert, M. J., & Andrews, A. A. (1996). How much therapy is really enough? A session-by-session analysis of the psychotherapy dose-effect relationship. *The Journal of Psychotherapy Practice and Research, 5*, 132–151. PMCID: PMC3330412

Kivlighan, D. M., Lin, Y. J., Egan, K. P., Pickett, T., & Goldberg, S. B. (2019). A further investigation of the good-enough level model across outcome domains and termination status. *Psychotherapy, 56,* 309-317. doi:10.1037/pst0000197

Kopta, S. M., & Lowry, J. L. (2002). Psychometric evaluation of the Behavioral Health Questionnaire-20: A brief instrument for assessing global mental health and the three

phases of psychotherapy outcome. *Psychotherapy Research, 12*, 413–426. https://doi.org/10.1093/ptr/12.4.413

Lambert, M. J., Burlingaume, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy, 3*, 249–258. https://doi.org/10.1002/(SICI)1099-0879(199612)3:4<249::AID-CPP106>3.0.CO;2-S

Lorenzo-Luaces, L., DeRubeis, R. J., van Straten, A., & Tiemens, B. (2017). A prognostic index (PI) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models. *Journal of Affective Disorders, 213*, 78-85. doi:10.1016/j.jad.2017.02.010

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*, e1000097. https://doi.org/10.1371/journal.pmed.1000097

Mulhall, D. (1976). Systematic self-assessment by PQRST. *Psychological Medicine, 6,* 591-597. PMID: 1005576

*Nielsen, S. L., Bailey, R. J., Nielsen, D. L., & Pedersen, T. R. (2016). Dose response and the shape of change. In S. Maltzman (Ed.), *The Oxford handbook of treatment processes and outcomes in psychology: A multidisciplinary, biopsychosocial approach* (pp.465-496). Oxford, UK: Oxford University Press.

*Owen, J. J., Adelson, J., Budge, S., Kopta, S. M., & Reese, R. J. (2016). Good-enough level and dose-effect models: Variation among outcomes and therapists. *Psychotherapy Research*, *26*, 22–30. https://doi.org/10.1080/10503307.2014.966346

*Owen, J., Adelson, J., Budge, S., Wampold, B., Kopta, M., Minami, T., & Miller, S. (2015). Trajectories of change in psychotherapy. *Journal of Clinical Psychology, 71*, 817-827. https://doi.org/10.1002/jclp.22191

*Reese, R. J., Toland, M. D., & Hopkins, N. B. (2011). Replicating and extending the good-enough level model of change: Considering session frequency. *Psychotherapy Research*, *21*, 608–619. https://doi.org/10.1080/10503307.2011.598580

Robinson, L., Delgadillo, J. & Kellett, S. (2019). The dose-response effect in routinely delivered psychological therapies: A systematic review. *Psychotherapy Research, 30,* 79-96. https://doi.org/10.1080/10503307.2019.1566676

Shapiro, M. B. (1961). A method of measuring psychological changes specific to the individual psychiatric patient. *British Journal of Medical Psychology, 34,* 151 – 155. https://doi.org/10.1111/j.2044-8341.1961.tb00940.x

*Stiles, W. B., Barkham, M., Connell, J., & Mellor-Clark, J. (2008). Responsive regulation of treatment duration in routine practice in United Kingdom primary care settings: Replication in a larger sample. *Journal of Consulting and Clinical Psychology, 76*, 298-305. doi:10.1037/0022-006X.76.2.298

*Stiles, W. B., Barkham, M., & Wheeler, S. (2015). Duration of psychological therapy: Relation to recovery and improvement rates in UK routine practice. *British Journal of Psychiatry, 207*, 115-122. https://dx.doi.org/10.1192/bjp.bp.114.145565

Stiles, W. B., Honos-Webb, L., & Surko, M. (1998). Responsiveness in psychotherapy. *Clinical Psychology: Science and Practice*, *5*, 439-458. https://doi.org/10.1111/j.1468-2850.1998.tb00166.x

*Stulz, N., Lutz, W., Kopta, S. M., Minami, T., & Saunders, S. M. (2013). Dose-effect relationship in routine outpatient psychotherapy: Does treatment duration matter? *Journal of Counseling Psychology*, *60*, 593–600. https://doi.org/10.1037/a0033589

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*, 419-435. http://dx.doi.org/10.1007/BF02294384