UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata

Corso di Laurea Magistrale in Psicologia Clinico-Dinamica

Tesi di laurea Magistrale

Bridging the gap between psychotherapy research and clinical practice: a multilevel meta-analysis of SCEDs, employing the MultiSCED software

Relatore

Prof. Enrico Benelli

Laureanda

Lorena Pizzocri

2048496

Anno Accademico 2022/2023

ABSTRACT

In the current era, scientific evidence in psychotherapy outcome research comes mainly from group studies - such as randomized control trials - that rely on comparing, on average, the effects of a treatment between two different groups. However, the ultimate focus of research regarding mental health issues should be the single individual and their intrapersonal changes, which can be hardly captured by between-person designs. Recently, the scientific community has been concerned by the inadequacy of group findings in applied settings (e.g., the clinical practice). Despite being able to prove high internal validity, they lack external validity and generalization, besides marginalizing a lot of emerging psychotherapy orientations due to their high costs.

A solution to this matter can be found by looking back at the origins of the psychotherapy history, when Freud and Breuer first published their *Studies on Hysteria*: single case designs. In fact, not only psychotherapy, but foundations of psychology, in general, sprouted from the study of individuals; some prestigious examples are Pavlov, Skinner, Watson, Broca, Alzheimer and Piaget.

Since those early years, single case experimental designs have evolved in many ways and now they can boast of rigorous statical methodologies. Their main strengths stand in quantitatively assessing the intervention effects, over time, within the same participant and being able to be integrated with qualitative information. For this reason, single case designs represent a way to bridge the gap between psychotherapy research and clinical practice, complementing the more widespread RCTs.

The present work aims to move a step in this direction. Indeed, a multilevel meta-analysis of 13 single case experimental designs (SCEDs) was conducted to evaluate the effects of Transactional Analysis (TA) psychotherapy on anxiety disorders, in an adult population. The Hierarchical Linear Model (a specific type of the multilevel modeling) was chosen with the exploratory aim to test its applicability in the psychotherapy outcome research field to assess treatments for common mental disorders.

To this end, the MultiSCED software was employed, as an innovative instrument that could help psychotherapy research to bridge the gap between research and clinical practice. Results need to be interpreted cautiously: while the majority of patients showed a reliable clinical change, only four of them presented significant regression coefficients at the individual level; besides, across all cases, none of the estimates did reach the statistical significance. Therefore, based on the current findings, it was not possible to affirm that the changes detected can be attributed to the effects of Transactional Analysis on treating anxiety disorders. Finally, limitations of the study and recommendations for future multilevel meta-analysis of SCEDs in this population are discussed.

Index

1.1 Psychotherapy Outcome Research: how treatments are supported and evaluated?

1.1.1 Empirically Supported Treatments (EST)

Three decades ago, in 1993, the Task Force on Promotion and Dissemination of Psychological Procedures [APA] published the first official criteria for identifying "*empirically validated psychological treatments*", also known today as "***empirically supported treatments*** **(EST)***"* (Chambless & Hollon, 1998; Chambless & Ollendick, 2001). Back at that time, the aim was the validation of specific procedures for each psychological problem and the spread of the results to help mental health professional and stakeholders to choose effective treatments for their clients.

The Task Force's work resulted in the institution of two sets of criteria to determine whether a treatment is "*well established*" or "*probably efficacious*" (Chambless & Hollon, 1998; Chambless & Ollendick, 2001). In the first case, a "*well established"* treatment must be supported by (a) at least two independently conducted, well-designed studies or (b) a large series of well-designed and carefully controlled single-case design experiments. In the second case, a *"probably efficacious"* one must be sustained by at least one well-designed study or a small series of single-case design experiments (Chambless & Hollon, 1998) . In the current research language, well-designed studies are referred as *randomized control trials (RTCs)* while single-case designs are referred as *single-case experimental designs (SCEDs).*

In 2006, following the EST research stream, a Presidential Task Force of the American Psychological Association adapted the medical classification of evidence-based medicine (EBM) to the mental health field, defining ***evidence-based practice*** **(EBP)** as the integration of three sources of information: (a) the best available research evidence on a treatment *efficacy* and *effectiveness*, (b) clinical expertise and (c) client characteristics (Goodheart et al., 2006). Evidence-based psychotherapy includes a wide set of procedures such as assessment, case formulation, therapeutic alliance factors and treatment decisions that guide the clinician to achieve the best possible outcome with the patient (APA Presidential Task Force on Evidence-Based Practice, 2006).

During the last two decades, ***Randomized control trials*** (**RCTs)** became the *gold standard* for supporting *efficacy* of the evidence-based treatments thanks to their intrinsic statistical characteristics of being able to prove high internal validity. The link between EBP and RCTs has become so rooted in the psychotherapy research field that, as Gold mentioned in 2015 "it almost seems necessary to remind readers that RCTs were not invented by it".

The core rationale behind the trials' approach consists in comparing the effects of an intervention on a treated group with the effects on a control group which either does not receive any treatment or receives a different one. The RCTs' ultimate methodology relies on: (a) pre-registration of the trial to mitigate biases in favor of publishing successful outcomes only; (b) random allocation of participants into various experimental conditions; (c) standardized treatments to ensure consistency across different practitioners administering the interventions; (d) utilization of intention-to-treat (ITT) analysis to counteract biases caused by participant dropouts; and (e) the ability to conduct meta-analyses and accurately determine effect sizes.

However, the efficacy of a treatment can be supported not only by RCTs but, also by a series of Single-Case Experimental Designs (SCEDs) with systematic replication by independent research groups (Chambless & Hollon, 1998). Despite this, over the past 25 years, the SCEDs' methodology has been neglected, resulting in an exponential grow of larger and more complicated RCTs at the expenses of those psychotherapy orientations that are not able to afford their high costs.

At this point, the reader could wonder why we are making such a case against the vast use of randomized control trials in psychotherapy research. In the following chapters, the advantages and the drawbacks of this approach will be presented, while proposing a valuable complementary approach to prove the efficacy of psychotherapy treatments: the **meta-analysis of single case experimental designs (SCEDs)**. For this purpose, a **Multilevel meta-analysis** of 13 SCEDs will be conducted, employing the innovative web software "**MULTISCED**" (Declercq et al., 2020).

1.1.2 Randomized control trials: efficacy comes at a cost

The need of psychotherapy research for quantitative methods comes from the 1950's. During that period, enormous efforts were made to demonstrate the efficacy of psychotherapy interventions to answer back Eysenck and his provocative "meta-analysis" stating that there wasn't any substantial evidence for psychological treatments. In this context, stringent numerical methods were made necessary to gain an acknowledged position in the scientific field and society in general (Braakmann, 2015). After this accomplishment, in the 1970's, the first RCTs were introduced from pharmacological research (where the trials were used to evaluate the impact of drug treatment) as the most up to date methodology: the aim was comparing several emerging therapeutic models to find out which were potentially more effective than others (Desmet, 2013).

Since then, for more than 50 years, randomized control trials have been the foundation for establishing whether a treatment is evidence-based and the basic condition to take part in the *horse race* among psychotherapy's orientations.

When an instrument is taken as the *gold standard* within the scientific field, it is a common praxis to keep using it (and continuously upgrading it) without ever questioning the historical moment and purpose it was created for. Indeed, it is very likely that the motivations that kickstarted the adoption of randomized control trials more than five decades ago are not the same that are driving psychotherapy research in the current era and that this method should be revised, considering the substantial amount of arisen criticisms.

According to the APA Task Force, the primary property to provide a treatment with the evidence-based practice definition regards the demonstration of its *efficacy* and *effectiveness* through the best available research. Secondly, clinical expertise and client characteristics have to be taken into account (APA Presidential Task Force on Evidence-Based Practice, 2006).

The glossary of Cochrane Collaboration (The Cochrane Collaboration 2005) defines **efficacy** as "The extent to which an intervention produces a beneficial result under ideal conditions" while

effectiveness as "The extent to which a specific intervention, when used under ordinary circumstances, does what it is intended to do". On one hand, the goal of efficacy is reached by research designs under systematically controlled conditions with the aim of accomplishing high *internal validity*. On the other hand, effectiveness is investigated studying the impact of treatments under "natural" conditions (the clinical setting), with *external validity* and *generalization* being the most important quality marker (Lambert & Ogles, 2004).

Nowadays, the scientific community is concerned by the huge gap between research and practice being caused by randomized control trials (Braakmann, 2015). On the one hand, RCTs employed to support psychological treatments in the public policy aim to maximize internal validity, which requires using a strict experimental methodology that eliminates alternative explanations in order to establish a causal relationship between the independent variable (treatment) and dependent variable (outcome) (Philips & Falkenström, 2021). On the other hand, this crave for creating such stringent experimental conditions often leads to overlook the problem of **effectiveness** in the daily routine: neither patient characteristics nor clinical expertise are considered, automatically violating two out of three assumptions for guaranteeing the evidence-based practice status. Here are some examples.

To conduct such designs, homogeneous group of patients should be selected with the aim to decrease sample variability and obtain a strong test of intervention effect (Kazdin, 2021); therefore, researchers often inspire to collect patients with isolated symptoms and co-morbidity is avoided as far as possible. It becomes immediately clear how this choice creates a discrepancy with the clinical practice, where such patients are extremely rare (Westen et al., 2004). In 2011, Westen and Morrison published a meta-analysis on the exclusion rates in RCTs, showing that approximately 65% of people in the patient group are routinely excluded because Axis I co-morbidity.

Furthermore, RCTs typically focus on the assessment of symptoms reduction without taking into consideration any other domain - such as psychosocial functioning (e.g., attendance to work and

performance), social engagement or family functioning - that Tolin et. al. (2015) defined as the *sine qua non* of a satisfactory treatment outcome.

Finally, the treatments provided in randomized control trials are strictly manualized and only a limited number of techniques are offered. Patients' preferences are not considered although they have been found to have a remarkable impact on the outcome (Kocsis et al., 2009; Raue et al., 2009). Along with patient's characteristics, in RCTs, also all the clinical expertise is totally discounted, given that professionals must follow strict manualized procedures.

Another relevant criticism about randomized control trials regards their apparent **generalizability**. There's a widespread belief that such group studies, with a great number of participants and a rigorous statistical methodology, are much more likely to yield results that can claim objectivity in procedures and, therefore, are inevitably *generalizable* to the entire population (Kazdin, 2021; Truijens, 2017). Nowadays, several research demonstrate that RCTs present some ambiguous statistical assumptions (e.g., Shean, 2012; Westen et al., 2004).

Firstly, as mentioned above, homogeneous group of patients should be selected with the aim of decreasing sample variability and obtaining a strong test of intervention effect. Apart from causing low effectiveness, this raises some critical issues regarding generalizability of results. Participants in psychotherapy between-group research are hardly sampled in a randomized way from the whole population of clinical patients: while assignment of participants to groups is random, selection of the sample from the population is not (Kazdin, 2021).

Kazdin (2021) also underlined that results in RCTs are analyzed in such a way – typically comparing means among groups- that does not tell anything about *how many* individuals in the group showed a real change in their symptoms. Given that typically mental health issues (e.g., anxiety or depressive disorders) are intrapersonal, treatment effects happen over time *within* individuals, therefore between-person comparisons may not always capture them (Maric et al., 2012; Schuurman, 2023).

Furthermore, the statistical rationale behind RCTs was adopted from the medical world without considering that psychological aspects of it– such as the randomization of patients across groups and the placebo effect – are critically challenging for psychotherapy research (Desmet, 2013). Double-blind administration of the treatment is impossible since the therapist obviously knows whether a real psychotherapy or a control is being provided. This knowledge interferes with the professional's expectations towards the treatment outcome, making the measurement of efficacy biased. Additionally, the use of a control group is problematic: following the pharmacological assumption to control for a possible psychological effect of the treatment, this makes no sense in psychotherapy research, since all effects are psychological (Desmet, 2013).

Desmet (2013) also pointed out that the *reliability* of RCTs is influenced by the length of the treatment: the shorter it is, the higher internal validity of the findings will be obtained. This is due to the presence of possible confounding variables – such as spontaneous recovery or personal life events – that can influence the patient's recovery path produced by the treatment itself. The treatment length of a RCT typically goes from 6 to 20 sessions, which is significantly shorter than psychotherapies in the widespread clinical practice (DeFife et al., 2015). Apart from the fact that is ridiculously illogical to adapt therapy to align with a research design, rather than accommodating the research design to suit the therapy, this raises a lot of concerns about whether the patient's clinical change promoted by the EBP movement is reliable.

In 1994, Ilardi and Graighaid showed that the first five sessions of therapy foster a lot of improvement, even though during this time, the intervention is barely administered. Other researchers even demonstrated that there is a significant improvement just after making the initial phone call to schedule the appointment for the first therapy session (Kopta et al., 1994 in Westen et al., 2004). This kind of progresses, however, are typically short term and have been interpreted as being the result of the restoration of hope in the patient (Howard et al., 1993). In addition, follow-up studies revealed that up to 88% of patients that terminate the treatment because of these first improvements, seek further therapy within the following two years (Westen et al., 2004). Undoubtedly,

psychotherapy outcome research is finding itself in front of a paradox: in the attempt of preserving the internal reliability of RCTs, it doesn't consider their external validity and, consequently, treatments delivered on the market as result of the best available research, often lack of a reliable clinical change over time.

Finally, the common practice of adopting short research designs, such as RCTs, fosters another important concern: *marginalization.* Proponents of all approaches face the pressure to conduct a greater number of high-quality randomized controlled trials in order to meet the requirements of evidence-based practice. Approaches that lack evidence from RCTs are marginalized and excluded, while those with extensive evidence, like CBT, receive more attention. This ideology rooted in evidence-based practice tends to favor treatments that are easier to study and show immediate after-treatment results. For instance, these treatments may involve fewer sessions, can be standardized, delivered efficiently in groups, or have simpler objectives focused on symptom change rather broader functional improvements (Stiles et al., 2015). It is noteworthy to mention that conducting these trials entails high costs and, therefore, all the psychotherapy orientations (e.g., emergent or innovative methods) that can't afford them get automatically cut off.

As a consequence of the EBP system, the rankings on various lists tend to sustain themselves, as approaches, lacking evidence from RCTs, struggle to gain meaningful representation on grant review committees and guideline development groups (Stiles et al., 2015).

For the above-mentioned reasons all the psychotherapies that can't conduct RCTs and don't gain the EST status are grouped under the label of Marginalized and Emerging Psychotherapies (MEPs).

1.1.3 IAPT program: application of evidence-based treatments by public health policies

What are the advantages and the consequences coming from the adoption, in the mental health policies, of evidence-based treatments exclusively supported by RCTs? In United Kingdom, the "Improving Access to Psychological Therapies (IAPT)" program could be taken as an example. The latter was introduced in 2008, with the aim to provide free evidence-based psychological treatments for anxiety and depression all over the country. It counts over 200 IAPT services across England and receives around 1.25 million annual referrals, qualifying itself as "*the largest publicly funded and systematic implementation of evidence-based psychological care in the World"* (Wakefield et al., 2021, p. 2).

The treatments offered follows the guidelines of the National Institute for Health and Care Excellence (NICE) with recommended evidence-based psychological interventions for common mental health disorders organized in a stepped care model (Bower & Gilbody, 2005). The latter is an evidence-based procedure, supported uniquely by controlled trials (Firth et al., 2015), in which progressively intense psychological treatments are delivered to patients according to symptom severity; the interventions range from guided self-help CBT (< 8 sessions) to high-intensity psychological therapies (typically 16-20 sessions). In this case, qualified therapists follow evidence-based protocols designed mainly from cognitive behavioral therapy, but also from other orientations such as interpersonal psychotherapy (IPT), eye movement desensitization and reprocessing (EMDR), brief dynamic interpersonal therapy (DIT) and couples counselling (Wakefield et al., 2021).

The reports of IATP program are published, monthly, on the official National Health service UK website and made available to the population. According to the data of the annual report 2021-22, out of the 1.24 million referrals accessed to IATP, only 50.2% moved to recovery and the average sessions completed were 7.9 (Population Health et al., 2022). These numbers are satisfactory but, as already Wakefield et al. (2021) pointed out, they leave room for improvement. To avoid any misunderstanding: we don't mean to question the efforts of UK in delivering free access to psychotherapy, which are virtuous and should be taken as a model for other developed countries, but

we are reporting some observations coming from the solely adoption of RCTs in setting guidelines for the clinical practice.

First of all, it comes natural to question where the remaining 50% that didn't benefit from recovery ended up. Secondly, the standardized protocols don't seem to work for patients with complex presentations, which are around 30% of who access to the IATP program (Delgadillo et al., 2017).

Last but not least, there is lack of studies with solid post-treatment follow-up data, thus the durability of IAPT interventions is still questionable (Wakefield et al., 2021). Also considering that – being the average sessions completed 7.9 – a lot of patients don't even reach the help of a professional and go through self-help interventions, how long do the effects of these short treatments last? As mentioned in the previous paragraph, the improvements resulting from the first psychotherapy sessions are often transitory and can't be considered as indicators of long-term change (Howard et al., 1993; Westen et al., 2004a).

Therefore, it is undeniable that the data presented by the National Health service of United Kingdom about the effectiveness of the IATP program are vague and leave room for improvement. It looks like that the adoption of a stepped care intervention, which can boast of the most cutting-hedge research methodology (RCTs), is missing something, namely that big portion of population that can't benefit from this approach. The IAPT program shows that, today, more than ever, it is necessary to accompany controlled trials with alternative research methods that allow to take into account both the complexity of each patient and of the clinical setting.

Single case experimental designs (SCEDs) can be a valid candidate to this aim.

1.1.4 Recommendations for EST updated guidelines

In 2015, Tolin et al. introduced some recommendations for a revised set of criteria for *empirically supported treatments* (EST). Indeed, the advancements in research quality and the several critics arose over the past two decades lead the authors to claim that the old criteria by Chambless and Hollon (1998) were outdated. Table 1 reports the complete list of critiques and proposals from Tolin et. al (2015): for our purposes, we are going to focus on the most salient ones.

Area	Critiques	Proposed change
Concerns about the strength of treatment	•Inadequate attention to null or negative findings **•Reliance on statistical, rather than clinical, significance** **•Inadequate attention to long-term outcomes** •Potentially significant variability in study quality	•Emphasize systematic reviews rather than individual studies •Separate strength of effect from strength of evidence **•Grade quality of studies** **•Consider clinical significance in addition to statistical significance** **•Consider long-term efficacy in addition to short-term efficacy**
Concerns about selecting among multiple treatment options	•Within a given EST category, there is little basis for choosing one over another •Lack of clarity about whether empirical support translates to a recommendation	• Present quantitative information about treatment strength • Make specific recommendations based on clinical outcomes and the quality of the available research
Concerns about the relevance of findings	**•Inadequate attention to functional outcomes** **•Inadequate attention to effectiveness in non-research settings or with diverse populations**	**•Include functional or other health-related outcomes as well as symptom outcomes** **•Address generalization of research findings to non-research settings and diverse population**
Concern about unclear active treatment ingredients and the proliferation of manuals for specific diagnoses	• Listing of packaged treatments rather than empirically supported principles of change •Emphasis on specific psychiatric diagnoses	•Evaluate and encourage dismantling research to identify empirically supported principles of change •De-emphasize diagnoses and emphasize syndromes/mechanisms of psychopathology

Table 1. *Common critiques of the EST movement and suggested changes* (Tolin et al., 2015)

First off, the authors expressed their concerns about the extreme reliance of RCTs on statistical rather than clinical significance, highlighting many issues such as inability to deal with complex patients, low generalizability, scarce effectiveness, symptom reduction as the only metric of improvement and insufficient attention to long-term outcomes (Tolin et al., 2015).

The logical path they adopted to propose changes is the result of the integration of the EST perspective with the APA evidence-based treatment guidelines (APA, 2006): EBP can be viewed as an approach to empirically supported treatments (ESTs), wherein the scientific information – coming from the most cutting-edge research – is filtered through the perspectives of both the clinician and the patient (Tolin et al., 2015). Therefore, in response to the critics reported above, recommendations for new criteria stressed on the importance of assessing a treatment also in terms of *effectiveness* in *non-research settings,* considering more *diagnostically complex patients* and preferring *open-ended*, *flexible practice* instead of manualization. Symptoms reduction must not longer be the only marker for efficacy since its value is hugely diminished if *functional improvements* (such as work attendance or performance, social engagement, and family functioning) are not equally demonstrated (Tolin et al., 2015).

Furthermore, at the end of their dissertation, Tolin and colleagues (2015) employed a modified version of the GRADE system by Guyatt et al. (2008), to rate the evidence quality of psychological treatments. According to this classification, the highest level of recommendation is given when there is high-quality evidence that the treatment has a clinically meaningful effect both on the symptoms and on functional outcomes, with significant improvement observed at the immediate post-treatment stage and at a follow-up interval of at least three months, with a small probability of harm and appropriate use of resources, and when there is, at least, one well-conducted study that shows the effectiveness of these findings in non-research setting.

Single case experimental designs (SCEDs), already appointed in 1998 by Chambless and Hollon as a valuable way to demonstrate efficacy and effectiveness of treatments but, unluckily,

neglected for many years in the psychotherapy research panorama, are up to the task today more than ever, being able to accomplish all the requirements of the GRADE system (Guyatt et al., 2008).

Unsurprisingly, Tolin et. al (2015) included **systematic reviews** and **meta-analysis of single case experimental designs (SCEDs)** their recommendations for a revised set of criteria for ESTs, supporting that, when using appropriate experimental control, they can establish causality in a manner comparable to RCTs (Shadish, 2014; Declercq et al., 2020). However, they suggest to not solely base on evidence from single-subject designs but accompanying them with larger clinical trials.

1.2 Meta-analysis of SCEDs: bridging the gap between research and clinical practice

1.2.1 Single case experimental designs to empirically support treatments: characteristics

Before diving into the topic of meta-analysis of single cases, we will focus on its main unit: SCEDs; we will provide a definition and outline their main features.

The scientific community seems to agree that the essential feature of single-case designs is that they require ongoing assessment of the intervention effects replicated within the same participant(s) over time (Kazdin, 2021; Smith, 2012). Indeed, individual cases (e.g., subjects) are regularly measured during a baseline condition which is followed by a treatment condition (Kazdin, 2011; Kratochwill & Levin, 2014). Patients are used as his or her own control to draw conclusions and the efficacy of a treatment is evaluated through repeated observations of the client's symptoms or functional changes over time (e.g., in the treatment of clinical outpatients, self-report is particularly adequate)(Kazdin, 2019). Researchers can determine whether there is an effect of the intervention on the outcome variable by comparing the scores obtained under the baseline condition with the ones resulting from the intervention condition. If the pattern of the intervention differs from the baseline, significant change has occurred.

Indeed, a fundamental distinction between group research and SCEDs is the importance of continuous assessment in the latter. In group studies, one group gets therapy while the control group does not. One or two observations (pre- and post-treatment assessments) are typically collected for every participant in each group to answer the problem of whether treatment results in change. In single-case studies, the effects of the treatment can be investigated by comparing how the same person performs with and without it. Multiple observations are gathered for one or a small number of people as opposed to one or two observations of multiple people, as in group research. At the end of the day, between-group and single-case designs follow the same rationale: making causal conclusions about the effects of treatments, allowing comparisons of performance under various contexts (Kazdin, 2019).

In single-case research, several designs are available (e.g., AB designs, Multiple Baseline designs, Changing-Criterion designs, etc.), each of which works by developing and testing hypotheses in various ways in accordance with patient's functioning (the data). What makes such designs particularly suitable for the clinical practice is their flexibility to adapt to the fluctuations of the individualized care and to easily adjust to changes (Kazdin, 2019). Indeed, SCEDs' logic puts the person in the center and values each patient's unique features, in contrast with group designs where the person is being sacrificed in order to adhere to rigid statistical rules. The most evident example of this lies in the number of sessions considered by those two research methods. While results of RCTs typically rely on studies with treatment conditions having a fixed number of sessions (usually between 6 and 20), single case designs are suitable for assessing treatments of any length, and, consequently, they result to be way more adequate to adjust to real clinical setting and to test the *long-term efficacy* of the treatment in question. This is just one example of how SCEDs hold the potential to address the concerns raised by both academics and clinicians regarding the *external validity* of controlled trials in the clinical practice and reported by Tolin et. al (2015) in their recommendations for EST updated guidelines. In the following paragraphs, other important points will be discussed.

Firstly, single cases can be conducted directly in the real clinical population, without the need to select a homogeneous group of patients with isolated symptoms and no co-morbidity. This allows the results of SCEDs to be more *generalizable* than RCTs in the clinical population.

Secondly, single cases permit to consider patient and therapist characteristics, adapting both to the therapist's clinical expertise (which is often discounted in manualized trials) and the patient's treatment intentions. On top of this, in single case designs, not only symptoms reduction is regarded but higher attention is paid also to the *psychosocial functioning* of each patient (e.g., work/school performance). All together these points find a solution to the "concerns about the relevance of the findings" raised by Tolin et al. (2015) in their recommendations for EST updated guidelines.

The authors also highlighted the importance of considering clinical significance above statistical significance. As mentioned earlier, between-person comparisons may not always capture

intra-personal changes which are the ultimate focus of the research regarding the efficacy of treatments of mental health issues (Maric et al., 2012; Schuurman, 2023). Citing John Grimley Evans, gerontologist and researcher: "Health care managers and trialists may be happy for treatments to work on average; patient's doctors -in our case psychotherapists- expect to do better than that." (Evans, 1995, p. 462). Fortunately, SCEDs represent a valid complementary approach to RCTs in order to assess clinical significance, which is what, at the end of the day, makes the difference in the mental health care work. The *reliable change index* (RCI) by Jacobson & Truax (1991) is a valid tool to evaluate whether a patient is experiencing clinically significant change and it will be employed later in this study.

Specifically, in the SCEDs scenario, another method exits to assess a person's improvements: visual inspection. "*Visual inspection refers to reaching a judgment about the reliability or consistency of intervention effects by visually examining the graphed data"* (Kazdin, 2019, p.11). Without reporting in detail the complete method for this technique, it is worth mentioning that the latter is widely employed in the single case research and many authors consider it as one of the principal ways to evaluate the effects of an intervention (e.g., Hornstra et al., 2023; Kazdin, 2019; Wolfe et al., 2019). Rich information can be learned from the graphical representation of data, such as how unsteady the baseline phase was, whether a significant change occurred, and how constant it was over time. These information gets lost in controlled trials where only the mean scores of the pre-, post-, and follow-up values are reported (Vlaeyen et al., 2020).

Parallelly to graphical representations, a new possibility arises: visualizing progress in real time during the treatment, hence having an *ongoing feedback* while the latter is still in effect (Kazdin, 2021). Indeed, research demonstrates that, on average, patients that receive a feedback during each psychotherapy session reports more statistically significant improvements compared to the ones not receiving it; moreover the majority of clients having ongoing feedbacks shows a reliable change at the end of the treatment (Reese et al., 2009)

Furthermore, one of the biggest potentials of single case designs in the psychotherapy field is their ability to inform both researchers and professionals about the individual. In fact, even though the aim of this research is to evaluate treatment from a quantitative perspective, it is important to make the reader aware that single cases hold the enormous potential to be integrated with *qualitative information* about the patient (this scenario goes under the "mixed methods" category). Qualitative research permits to study individuals and human experiences much more intensively than between-groups studies, often employing narratives and making use or not of statistical techniques to evaluate the content. This is the case of the Hermeneutic Single Case Design (HSCED) by Elliott (2002), which allows to collect rich and comprehensive information about a patient's therapy using multiple sources and measures. In fact, this methodology combines quantitative outcome measures with detailed qualitative information about the patient and the process, records of therapy sessions and a hermeneutic reasoning procedure, with different judges, to evaluate the causal role of therapy in generating the outcome (Elliott, 2002). Apart from representing an intelligent way to combine outcome with process in psychotherapy research, methods such as the HSCED have the remarkable prospect to enrich not only the scientific panorama, but, also, to provide practical tools to clinicians and students.

In this direction, in 2013, an online database the "*Single Case Archive"* (https://www.singlecasearchive.com) was created with the aim to bridge the gap between research and practice. Until now, 3.471 cases, from 175 peer-reviewed journals published between 1955 and 2019, are included in the database which is constantly accepting new cases. The Single Case Archive includes cases from various theoretical perspectives, discussing patients from various age groups, with various issues, undergoing a variety of psychotherapies, which are studied using different methodologies (Meganck et al., 2022). This initiative represents an immense resource to both researchers who want to synthesize findings across homogeneous sets of cases, clinicians who want

to learn about a treatment for a patient with a certain diagnosis and also students who have access to a constantly updated database with tons of first-hand clinical knowledge.

Last but not least, single case designs are extremely advantageous from a practical perspective, since they are relatively low-cost and they can be conducted on a much smaller scale than RCTs (Cawthorne et al., 2023). For this reason, they may be a starting point for evaluating new interventions before conducting more expensive and time-consuming RCTs: thus, they represent a strategic way to legitimize *marginalized and emerging psychotherapies* (MEPs) (Gaynor & Harris, 2008).

1.2.2 Meta-analysis of SCEDs: characteristics

The increasing employment of meta-analysis of single cases experimental designs (SCEDs) in the evidence-based practice panorama is extremely beneficial to address most of the issues reported in paragraph 1.1.2, complementing the more widespread randomized control trials (RCTs).

Nowadays, there is a shared consensus that, in addition to RCTs, we need more idiographic research approaches that are able to detect the changes of single patients. While meta-analyses of RCTs focus on the evidence at the sample level analyzing *between-person* effects, meta-analyses of single-case studies explore *within-person* changes throughout treatment, looking at the effect at case level. Indeed, single case research allows to evaluate the outcome for specific cases, for specific interventions and in specific contexts. Despite this idiosyncrasy, it is also possible predict a certain communality: if a certain effect is found to be effective in some cases, it may also be expected to be effective in other cases (Maric et al., 2023). Therefore, the intervention's overall effects across all cases and studies can be assessed through a meta-analysis, which thanks to its enhanced power and generalizability, represents the cutting-hedge methodology for scientific evidence in psychological sciences (Gold, 2015).

Nevertheless, SCEDs meta-analysis enables measurement of the degree to which an effect differs between studies and cases, i.e., the degree to which the effects discovered are *heterogeneous* and *generalizable*. If treatment results are heterogeneous, it is possible to determine whether we can account for this diversity by looking at the mediating and moderating effects of case and study variables. In contrast to group comparisons, single-case meta-analysis estimates treatment effects for each individual, providing the opportunity to explore moderating effects of case characteristics (Gaynor & Harris, 2008).

The outcomes of well conducted SCEDs can be compared to those of RCTs if they meet the following criteria: (a) reliable and valid outcome measures; (b) on going assessment of key outcome variables over time; (c) stability of the baseline before treatment; (d) solid effect of the intervention

on outcome variables supported by time-series analysis; and (e) repetition of the same pattern over numerous cases (Kazdin, 1981).

1.2.3 Multilevel modeling and the MultiSCED

In 2003, Van den Noortgate and Onghena first proposed the application of multilevel modeling for meta-analysing SCED data and, since then, this approach has expanded in several methodological works (e.g., Moeyaert et al., 2020; Rindskopf & Ferron, 2014).

Similar to traditional meta-analyses, multilevel meta-analysis constitutes a practical and valuable approach to systematically assess research evidence across primary studies exploring the same research issue (Glass, 1976). The model is particularly convenient for summarizing hierarchical structured data (or data with a nested configuration), such as single case experimental designs (Van Den Noortgate & Onghena, 2003). Indeed, while conventional meta-analysis combine effect sizes across studies, "Multilevel meta-analysis are able to summarize participant-specific effect sizes across cases and across studies" (Moeyaert, 2019, p.1). The word "Multilevel" itself entails that there are *higher levels* and *lower levels* of data; for instance, the current study has two levels of analysis: "One-level analysis" which corresponds to the analysis of the individual case and "Two-level analysis" which aims to investigate whether the treatment effects can be generalized across multiple cases within a study. In the context of SCEDs, it is also common a "Three-level analysis": cases across studies are combined together in order to further generalize the estimates of the treatment effect (Moeyaert et al., 2014).

Despite its promising application in numerous fields, the practical use of multilevel meta-analysis is limited, possibly due to the perception of its complexity and its relative novelty in behavioral sciences: a recent review by Jamshidi et al. (2018) showed that only the 17% of SCEDs meta-analysis employed the multilevel modeling (Moeyaert et al., 2020).

Consequently, in 2020, Declercq and colleagues developed the web application "MultiSCED" (available at https://ppw.kuleuven.be/single-case/MultiSCED), with the aim to help researchers to

apply the Multilevel Modeling to quantitatively summarize single-case experimental designs (SCEDs). Indeed, as it has been highlighted by Manolov & Moeyaert (2017) and Shadish (2014), there is often a discrepancy between the advancements in statistics and the practical implementation by researchers in the field of behavioral studies. For this reason, a free and easy-to-use software was made necessary to implement this new and sophisticated methodology. The application provides a point-and-click user interface (Fig. 1) that enables practitioners to use the freely available R software environment without the need for an extensive knowledge of the R syntax coding.

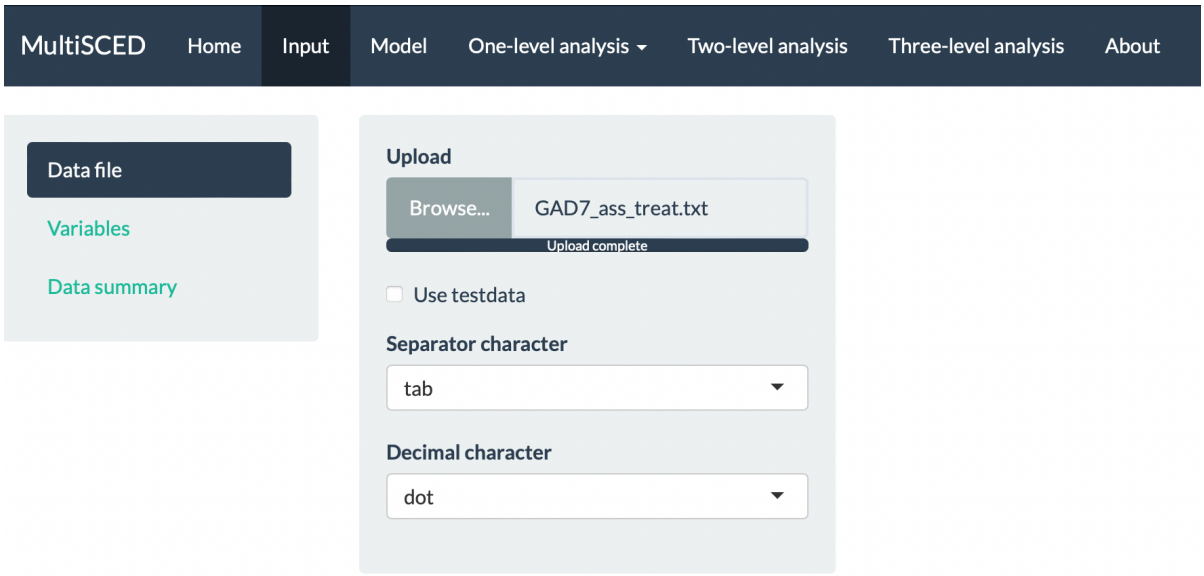


Figure 1. *Online interface of the MultiSCED*

In addition, before computing the analysis, the app displays the multilevel modeling equations (see chapter 3), giving users an accurate mathematical representation of the model they are using (Declercq et al., 2020). From a pragmatic perspective, MultiSCED aims to help users in understanding how to apply multilevel modeling to their SCED data analysis. In particular, MultiSCED allows to analyze data step by step, beginning with the standard (single level) linear regression model. After having explored the single-level data and regression results case by case, the app permits to merge them using a multilevel model (DeClercq et al., 2020).

In our study, a specific type of multilevel model has been employed: the **Hierarchical Linear Model (HLM)**. The latter implies a linear relation (represented by a straight line) between the predictor and the outcome variable. In the "Analysis" section, the equations of this model will be discussed more in detail.

1.3 Transactional Analysis

The chosen model of psychotherapy for the present study is Transactional Analysis (TA).

The latter was initially developed by the psychiatrist Eric Berne in the late 1950s both as a theory of personality, group dynamics and psychotherapy approach that places emphasis on an open and equal dialogue between the client and the therapist (Berne, 1958, 1961). For instance, TA deeply respects each person's objectives and drives creating explicit and mutually shared treatment goals known as "contracts". The center of the conceptual model of TA psychotherapy treatments is the ego-state model of personality, which is founded on the concept of ego states (Parent, Adult, and Child). While the Parent and the Child incorporate primitive experiences and introjects, the Adult is responsive to the present-day reality. Besides this, another foundation of TA is the therapeutic mechanism of Life position which explores the different attitudes and beliefs that people have about themselves and other, through the opposites "I am OK/not-OK" and "Others are OK/not-OK" (Vos & Van Rijn, 2021a, 2021b)

Based on these core principles, several distinct schools developed within the TA literature and practice such as: psychodynamic TA integrated psychodynamic theories and methods (Moiso & Novellino, 2000), cognitive-behavioral TA (Bergmann, 1981), co-creative TA (Summers & Tudor, 2018), relational TA integrated relational psychoanalysis (Hargaden & Sills, 2014),etc.

Although it is largely taught and practiced internationally within high-standard academic and professional institutions, TA is still not well established in the psychotherapy research area(Vos & Van Rijn, 2021a). In 2022, Vos and van Rijn published the results of a literature review and meta-analysis including all the studies about Transactional Analysis treatment realized so far, with the aim to evaluate the effectiveness for such model. The outcome indicates that, overall, TA accomplishes the general purpose of psychotherapy of providing symptoms relief and personality change, preventing the repetition of symptomatic episodes in the future, improving the quality of life, encouraging adaptive performance in job, school and relationships, and increasing the likelihood of maintaining a healthy life (APA Presidential Task Force on Evidence-Based Practice, 2006).

Despite this, Transactional Analysis is still working toward the achievement of the EST status and, consequently, it still struggles to be included in the standard policies by the mental health services and insurances. In other words, TA can be considered a *marginalized and emerging psychotherapy* (Benelli & Zanchetta, 2019; Vos & van Rijn, 2022). According to Vos and Van Rijn (2022), this could be because studies for each of the contexts of TA application are limited and most of them are relatively old, thus not following the updated guidelines from Chambless and Hollon (1998) and Tolin et al. (2015). For instance, in psychotherapy research literature, there is little and incomplete evidence regarding the outcome of Transactional Analysis for the treatment of anxiety disorders (e.g., Solgi et al. 2021), even though this model is successfully employed in the clinical practice and treatment manuals have been recently developed (Benelli et al., 2021).

1.4 Objectives of the study

For all the above-mentioned reasons, the present study has three main goals:

a) Filling gap in psychotherapy outcome research regarding the efficacy and effectiveness of Transactional Analysis in treating anxiety disorders. To our knowledge, neither a RCT nor a meta-analysis of SCEDs have been ever published on this topic until now. Through this study, we would like to move a step forward the recognition of TA as an *empirically supported treatment* according to the recent guidelines from Tolin et. al (2015).

b) To this purpose, a multilevel meta-analysis of 13 single case experimental designs (SCEDs) will be computed. This specific meta-analytic method has been chosen with the exploratory aim of testing its applicability in the psychotherapy outcome research field to assess treatments for Common Mental Disorders. Indeed, despite multilevel modeling meta-analysis of SCEDs are becoming more and more sophisticated from a methodological perspective (e.g., Baek et al., 2023; Moeyaert et al., 2020), there is almost no evidence in literature of the employment of this method for evaluating the efficacy of psychotherapy models in treating psychopathology (i.e., Maric et al., 2023).

c) Finally, the software MultiSCED by Declercq et al. (2020) has been employed for conducting the multilevel meta-analysis. This cutting-hedge web application provides a point-and-click user interface that enables practitioners to use the R environment without the need for an extensive knowledge of the R syntax coding. We believe that the MultiSCED represents an important opportunity to bridge the gap between research and clinical practice since psychotherapist, clinicians and researchers could benefit of an instrument that allows to rigorously test and summarize the effects of their clinical work.

2. Methods

2.1 Population

The data for the present research were collected through the program "*Assessment of efficacy and clinical effectiveness of emerging and marginalized models of psychotherapy by collecting an online database from clinicians and researchers through single-case designs*" promoted by the University of Padua, together with the "Instituto di Analisi Transazionale Psicodinamica (IATP)". Scores have been gathered by independent Italian clinicians or researchers (anonymous to the authors of this study) that agreed in participating in the current research.

In the original database there were 41 SCEDs. However, given that a substantial number of them presented important lacks (such as subclinical scores during the baseline phase or not enough measurements), they were excluded from the present research.

The inclusion criteria for participating in the current study were set as following:

- Age ≥ 18 years old
- Being diagnosed with an anxiety disorder according to the DSM-V
- Scoring in the clinical range (score ≥ 10) during the assessment phase on the *generalized anxiety disorder-7* scale (Spitzer et al., 2006);
- Total number of measurements per participant ≥ 18

Applying the above-mentioned conditions, the current meta-analysis comprehends a total number of 13 SCEDs. Most of the population is female (10 out of 13) and they are relatively young (8 out of 13 patients have an age within the 18-30 range, M=34,62). The most frequent diagnosis is *generalized anxiety disorder* (7 cases) followed by *unspecified anxiety disorder* (4 cases), *panic disorder* (1 case) and *social anxiety disorder* (1 case). Comorbidity with other common mental health disorders or personality diagnosis was admitted. The final number of cases, even though it is markedly lower than the initial sample, is line with another recent single case design meta-analysis supporting the TA treatment for depression (Benelli & Zanchetta, 2019).

2.2 Intervention

All clinicians involved in the present study followed the manualized therapy protocol for the treatment of anxiety disorders of Benelli et al. (2021). The therapists received supervision from a weekly to monthly basis from a Provisional Teaching and Supervising Analyst (Psychotherapy) (PTSTA-P), or from a Teaching and Supervising Transactional Analyst (Psychotherapy) (TSTA-P).

2.3 Procedure

Patients who accessed to the clinical/research centers involved in the present study were offered to take part in a research program with the aim of evaluating the effectiveness of *marginalized and emerging psychotherapies* (MEPs) for Common Mental Disorders. Participants were informed that the research protocol was about a long-established treatment but still not empirically supported (marginalized). They were also informed that participation involved completing some self-reports for 10 minutes, through an online platform, before each session. In particular, the research entailed: 3 assessment sessions, from 16 to 40 treatment sessions of manualized psychotherapy and three follow-up measurements. The assessment and treatment phase consisted in weekly sessions lasting one hour. Before joining the research, all the participants received detailed information regarding its purpose, the type and the length of the treatment protocol, the tests/questionnaires they had to fill in, and how their sensitive data would have been processed. If they agreed to take part in the research, they were asked to sign an informed consent. Participants were allowed to withdraw from the research at any time, without any consequences on the ongoing treatment.

Researchers did not get in touch with patients since they were solely in charge to monitor and analyze the data, previously collected in the clinical practice.

2.4 Measure

The **Generalized Anxiety Disorder-7** scale (Spitzer et al., 2006) was employed as the self-report measure to assess the presence and severity of anxiety symptoms throughout the baseline and the treatment phase. The scale measures the severity of symptoms associated with generalized anxiety disorder and shows a good internal reliability (Cronbach's alpha = .90; Spitzer et al., 2006).

In addition, it is also reasonably good at detecting three other common anxiety disorders: post-traumatic stress disorder (sensitivity 66%, specificity 81%), social anxiety disorder (sensitivity 72%, specificity 80%), and panic disorder (sensitivity 74%, specificity 81%) (Kroenke et al., 2007). Instructions ask: "Over the last 2 weeks, how often have you been bothered by the following problems?". Example items are: "Feeling nervous, anxious, or on edge" and "Being so restless that it's hard to sit still." Respondents answer on a 0 to 3 scale from "Not at all" to "Nearly every day". Scores range from 0 to 21, with higher scores indicating greater levels of anxiety: 0 to 4 indicates minimal anxiety, 5 to 9 mild anxiety, 10 to 14 moderate anxiety and 15 to 21 severe anxiety (Spitzer et al., 2006).

3. Analysis

Data were analyzed using a specific type of *multilevel regression model,* named **Hierarchical linear model (HLM).** As mentioned earlier, the latter implies a linear relation (represented by a straight line) between the predictor and the outcome variable. SCED data are analyzed within cases and combined across cases and studies.

To this end, the MultiSCED software (https://ppw.kuleuven.be/single-case/MultiSCED), a point-and-click web application built on the R syntax (see Appendix A), was employed (Declercq et al., 2020). Our data were particularly suitable for conducting a multilevel analysis since they met all the prerequisites indicated by Declercq et al. (2020) and reported below:

- Data need to be gathered according to an AB-phase designs: several measurements should be taken for each case throughout the baseline phase and treatment phase
- Data should include at least the following variables:
    - A variable indicating the participant.
    - A variable indicating the Phase (baseline vs treatment).
    - A variable about the measurement occasion (e.g., time or session).
    - An outcome variable: the value of the dependent variable for that specific measurement occasion.

Data were prepared according to the author's instructions, stored in the .txt format (see Appendix B) and uploaded onto the software where the variables "Author", "Name", "Phase", "Time" and "Y" (outcome) were selected. At this stage, we chose to center the time variable by clicking the option "Center time variable". This option transforms the time variable by setting the first observation from the treatment phase to zero. Finally, in the "Model" page, one level and two-level regression models were defined to conduct the analysis (Declercq et al., 2020).

In the "**One level analysis**", researchers can estimate the effects of case-specific treatments using the ordinary least squares regression analysis (Huitema & McKean, 1998). At this level, the independent variables are: Time_i – a variable indicating the measurement occasion (i.e., session), Phase_i – a variable indicating whether the measurement occasions are in the baseline or in the treatment phase – and their interaction (Time_i x Phase_i). Consequently, it is possible to define the one level model for each individual case (Declercq et al., 2020):

$$Y_i = \beta_0 + \beta_1 \text{Time}_i + \beta_2 \text{Phase}_i + \beta_3 (\text{Time}_i \text{ x } \text{Phase}_i) + e_i \qquad (1)$$

The measurement nested inside the case is indicated by the subscript *i.* The time at which measurement *i* was observed is indicated by the variable Time_i. It is assumed that the residuals e_i are independent and follow a normal distribution with zero mean and standard deviation δ_e. In this one-level model, the sampling error e_i is the only source of variation: this is the random variation of the sample measurements around the expected value. When a measurement Y_i is registered in the baseline phase, the variable score Phase_i of Eq.1 becomes zero. Hence, the single case model of the baseline data follows this simplified equation: a straight line with respect to Time, with an *intercept* β_0 and a *slope* β_1 (Declercq et al., 2020).

$$Y_i = \beta_0 + \beta_1 \text{Time}_i + e_i \qquad (2)$$

On the other hand, when a measurement Y_i is registered in the treatment phase, the variable score Phase_i of Eq.1 will be equal to one. Thus, the single case model of the treatment data has this equation:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \text{Time}_i + \beta_2 + \beta_3 \text{Time}_i + e_i \qquad (3) \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{Time}_i + e_i \end{aligned}$$

β_0+ β_2 is the *intercept* in the treatment phase (i.e., the predicted outcome score at the start of the treatment phase) and β_1+ β_3 is the *slope* for the treatment phase. Hence, β_2 is the effect of the treatment on the intercept, and β_3 is the effect of the treatment on the slope (Declercq et al., 2020).

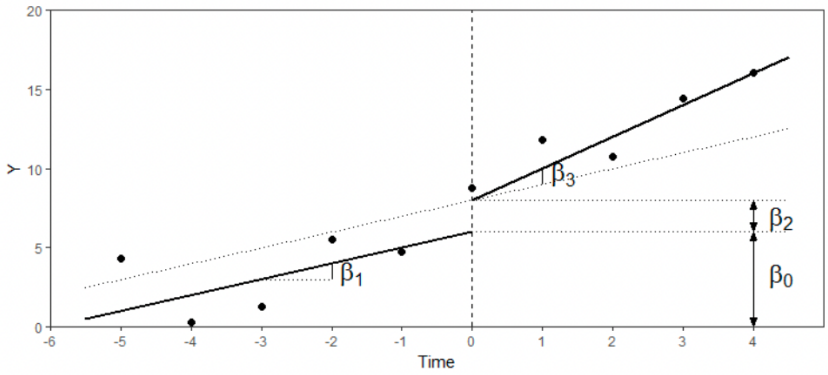


Figure 2*. Graphical interpretation of the one-level model parameters from Eq.1 after centering for the time variable (Declercq et al., 2020)*

At this level of analysis, the MultiSCED allows also to calculate the **effect size**. Within the multilevel model framework, the latter can be defined *as the outcome predicted by the treatment phase regression line minus the outcome predicted by the baseline phase regression line*.

In the current research, a *constant linear model without time trend* (Eq. 4) was employed, where the effect size prediction is simply equal to β_1.

$$Y_i = \beta_0 + \beta_1 \text{Phase}_i + e_i \qquad (4)$$

The goal of the "**Two level analysis**" is combining the single case data in order to investigate whether the treatment effects can be generalized across multiple cases within the same study (Moeyaert et al., 2014). Declercq et al. (2020) built further on the Eq.1, adding an index *j* to indicate case *j* with the study.

$$Y_{ij} = \beta_{0j} + \beta_{1j}\text{Time}_{ij} + \beta_{2j}\text{Phase}_{ij} + \beta_{3j}(\text{Time}_{ij} \text{ x } \text{Phase}_{ij}) + e_{ij} \qquad (5)$$

The standard deviation δ_e of the residuals e_{ij} (independent and normally distributed) is assumed to be identical for all cases. The present two-level analysis is a *fixed effects* meta-analysis (in contrast to *random effects* meta-analysis). Indeed, it has been assumed that there is a common treatment effect across all SCEDs and that any differences between the observed effect sizes are the results of sampling error (Nikolakopoulou et al., 2014).

Since more and more single-case studies have been published over the past ten years, there is a growing interest in meta-analyzing these studies to determine the average treatment effects.

The MultiSCED offers also the possibility to compute a "**Three-level analysis**" in order to combine data from various research. In the current study, this level of analysis was not employed given that the cases available came only from one single study. However, it is worth reporting the basics of this level to provide a complete illustration of the program potential. Using the "Three-level analysis" we can assess the generalizability of the findings by combining the results of multiple investigations. Therefore, important judgments might be taken based on these conclusions to inform public policies. To build the three-level model equation, Declercq and colleagues (2020) extended the two-level model Eq. 5 adding a new index k to indicate the study. At this stage the Eq. 6 includes measurements at the first level, cases at the second level and studies at the third level.

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}\text{Time}_{ijk} + \beta_{2jk}\text{Phase}_{ijk} + \beta_{3jk}(\text{Time}_{ijk} \text{ x } \text{Phase}_{ijk}) + e_{ijk} \quad (6)$$

On top of providing results in the numerical form, MultiSCED also displays them graphically, helping users in their interpretation. These graphs can be stored as figures to be used in data analysis reports to complement statistical analysis (see Figure 3 and 4). In the current study, the **visual inspection criteria proposed by Kazdin (2019)** were employed to evaluate the plots resulting from

one-level regression. In order to assess the magnitude and the rate of the changes across phases, the author advised to consider four main characteristics of the data (Kazdin, 2019):

- Change in means from phase to phase in the expected direction.
- Change in slope.
- Shift in level from one phase to another: change in level from the last day of one phase (e.g., baseline) and the first day of the next phase (e.g., intervention).
- Latency of change: the period between the termination of one phase (e.g., from baseline to intervention) and changes in performance. The closer in time change occurs after the conditions have been altered, the easier it is to attribute the change to the intervention.

Finally, the *Reliable Change Index* (RCI) for GAD-7 established by Bischoff et al. (2020) through the method of Jacobson & Truax (1991) was employed to assess whether the change of each single case was statistically different from a change due to random measurement error. Specifically, in the case of the GAD-7, a patient can be considered experiencing clinically significant change if their score crosses the cutoff of 10 and increases or decreases by 6 or more points between the first and most recent administration (Bischoff et al., 2020).

4. Results

4.1 One-level analysis: Within-Person Symptoms Change

The results of the one-level OSL regression for all the participants are presented in Table 2 and displayed graphically in Figure 3. Each row of the output in Table 2 represents a case-specific regression coefficient estimate, together with the standard error, *t-value* and *p-value*.

At the first glimpse, it can be observed that, in most of the cases, the standard error is particularly high, probably due to the important presence of outliers in our sample. This led to have only 4/13 cases with significant regression coefficients. In this section, only the subjects with significant regressors will be commented, while a more detailed analysis for each subject can be found in the Appendix C.

Subject 6 presents a very stable baseline, hence the estimated intercept at the beginning of the treatment phase is $\hat{\beta}_0$ = 16 ($p < 0.001$) and strongly significant. Since the patient scores the same during the entire baseline, the Time regression coefficient is equal to zero ($\hat{\beta}_1 = 0$), indicating that, without treatment, the person's scores relative to the symptoms remain the same (however, Time coefficient is not significant). According to the data, what does create a significant difference is the immediate effect of the intervention since, right after its start, a big drop in the patient's scores is registered ($\hat{\beta}_2 = -6.824, p < 0.05$). However, because of the non-significant *p-value* associated with the interaction between Phase1 and Time ($\hat{\beta}_3 = 0.076, p = 0.95$) , we can't affirm that that the treatment effect on the time trend is statistically significantly different from zero. This could be due to the influence of the outliers in the intervention phase.

This is not the case of subject 7, whose score at the beginning of the treatment phase would be $\hat{\beta}_0$=17.667 ($p<0.001$), together with the coefficient Time:Phase1 ($\hat{\beta}_3 = -2.449, p = 0.052$) being nearly significant. This indicates that, starting from the intercept $\hat{\beta}_0$=17.667, the intervention effect on the time trend generates a decrease of $\hat{\beta}_3 = -2.449$ for every session.

Subject 3 shows a steep slope already in the baseline condition, where the expected score decreases by -4 points per time unit ($\hat{\beta}_1 = -4\,, p < 0.05$). On the contrary, the treatment phase doesn't seem to have itself a big impact in decreasing the symptoms which, however, remain steady (around $Y_i = 4$) on the time trend, with a slope in the treatment phase being almost equal to zero ($\hat{\beta}_3 = 4\,, p < 0.05$).

Similarly, subject n 12 presents an important decrease in score during the baseline condition ($\hat{\beta}_1 = -2{,}5\,, p < 0.05$) to which it is added the contribution (unfortunately not significant) of the start of the treatment phase. The latter doesn't seem to have itself an important impact in decreasing the symptoms which, however, remain steady (around $Y_i = 2{,}5$) on the time trend, with a slope in the treatment phase being almost equal to zero ($\hat{\beta}_3 = 2{,}429\,, p = 0.052$).

***Table 2**. One level regression results; * = p < 0.05*

Subject	Regressor	Coefficient	Standard Error (SD)	t-value	p-value
1	(Intercept)	13	3.89	3.342	0.004
1	Time	-1	1.8	-0.555	0.587
1	Phase1	-0.412	4.075	-0.101	0.921
1	Time:Phase1	0.638	1.806	0.353	0.729
2	(Intercept)	5.667	2.577	2.199	0.044
2	Time	-2.5	1.193	-2.096	0.053
2	Phase1	2.341	2.7	0.867	0.4
2	Time:Phase1	2.107	1.196	1.762	0.099
3	(Intercept)	1	3.566	0.28	0.783
3	Time*	-4	1.651	-2.423	0.029
3	Phase1	3.375	3.736	0.903	0.381
3	Time:Phase1*	4	1.656	2.416	0.029
4	(Intercept)	12	3.465	3.463	0.003
4	Time	-1	1.604	-0.623	0.542
4	Phase1	-1.478	3.63	-0.407	0.69
4	Time:Phase1	0.972	1.609	0.604	0.555
5	(Intercept)	5	6.416	0.779	0.448
5	Time	-3	2.97	-1.01	0.328
5	Phase1	3.169	6.722	0.471	0.644
5	Time:Phase1	2.869	2.979	0.963	0.351
6	(Intercept)	16	2.587	6.186	0
6	Time	0	1.197	0	1
6	Phase1*	-6.824	2.71	-2.518	0.024
6	Time:Phase1	0.076	1.201	0.064	0.95

7	(Intercept)	17.667	2.501	7.064	0
7	Time	2	1.158	1.728	0.105
7	Phase1	-3.865	2.62	-1.475	0.161
7	Time:Phase1*	-2.449	1.161	-2.109	0.052
8	(Intercept)	3.333	3.21	1.038	0.316
8	Time	-2.5	1.486	-1.682	0.113
8	Phase1	3.328	3.363	0.99	0.338
8	Time:Phase1	2.387	1.49	1.601	0.13
9	(Intercept)	6.667	5.195	1.283	0.22
9	Time	-1.5	2.405	-0.624	0.543
9	Phase1	5.3	5.457	0.971	0.348
9	Time:Phase1	1	2.413	0.414	0.685
10	(Intercept)	14	4.9	2.857	0.012
10	Time	1.5	2.268	0.661	0.518
10	Phase1	-2.346	5.134	-0.457	0.654
10	Time:Phase1	-1.421	2.275	-0.624	0.542
11	(Intercept)	18.333	5.039	3.638	0.002
11	Time	3	2.333	1.286	0.218
11	Phase1	-3.319	5.28	-0.629	0.539
11	Time:Phase1	-3.435	2.34	-1.468	0.163
12	(Intercept)	2.667	2.374	1.123	0.279
12	Time*	-2.5	1.099	-2.275	0.038
12	Phase1	-0.762	2.487	-0.306	0.763
12	Time:Phase1*	2.429	1.102	2.204	0.044
13	(Intercept)	3.667	5.077	0.722	0.481
13	Time	-2.5	2.35	-1.064	0.304
13	Phase1	6.407	5.319	1.205	0.247
13	Time:Phase1	1.974	2.357	0.837	0.416

4.1.1 Effect size predictions

For each subject, the treatment *effect size* (ES) was calculated employing a constant linear model, following Eq. 4. This prediction is independent of time and it simply considers the coefficient β_1 from Eq.4, which is related to the change from Phase 0 (baseline condition) to Phase 1 (treatment condition) (Declercq et al., 2020). Results are reported in Table 3 under the column "ES" (MultiSCED also computes a confidence band $\alpha = 0.05$, see Appendix D for the complete table). All the subjects (except for subject n 10) present a negative effect size, ranging from - 6,29 to -0,58.

Subject	Y_0	Y_1	RC	ES	One-level
Subject 1	16	3	Yes	- 5,12	No
Subject 2	13	3	Yes	- 5,60	No
Subject 3*	13	4	Yes	- 4,62	Yes
Subject 4	15	10	No	- 3,69	No
Subject 5	14	4	Yes	- 3,81	No
Subject 6*	16	9	Yes	- 6,25	Yes
Subject 7	11	9	No	-3,23	Yes
Subject 8	11	6	No	-2,52	No
Subject 9	11	0	Yes	-1,20	No
Subject 10	10	10	No	1,25	No
Subject 11	10	4	Yes	- 0,58	No
Subject 12*	12	0	Yes	- 6,29	Yes
Subject 13	11	0	Yes	- 2,54	No

Table 3*. Comparison between reliable change index, effect size and results of the one-level OSL regression. * SCEDs that show both a reliable change, a negative effect size and significant coefficients in the one-level analysis.* Y_0 *= first measurement in the baseline phase.* Y_1 *= last measurement in the treatment phase. RC= has the patient experienced reliable clinical change? ES=effect size. One-level= at least one significant coefficient related to the treatment phase (cf. Table 2).*

4.1.2 Reliable change

Following the parameters established by Bischoff et al. (2020), it was calculated whether each participant experienced reliable change. The results are reported in Table 3 under the column "RC", together with the first GAD-7 measurement in the baseline phase and the last GAD-7 measurement in the treatment phase. Nine out of thirteen patients experienced clinical reliable change, subject n 10 did not seem to have experienced any change and both subjects 4 and 8 experienced some change but still not enough to be considered significant. Particular attention has to be paid to subject 7; according to the criteria by Bischoff et al. (2020), it seems that there is not a reliable clinical change, being the difference between first and last measurement not enough to claim so. In fact, the first assessment score is 11, the subsequent two are both 15 and 15 and the last assessment of the treatment phase is 9. On the other hand, this configuration of the data creates an increasing slope in the baseline phase, which differs from the decrease in slope during the intervention phase, generating a nearly significant coefficient on the Time:Phase1 regressor ($\hat{\beta}_3 = -2.449, p = 0.052$).

4.1.3 Visual inspection

In the following lines, the one-level analysis regression plots (Figure 3) will be presented employing the visual inspection suggestions provided by Kadzin (2019) and in the light of the statistical analysis of the previous paragraphs.

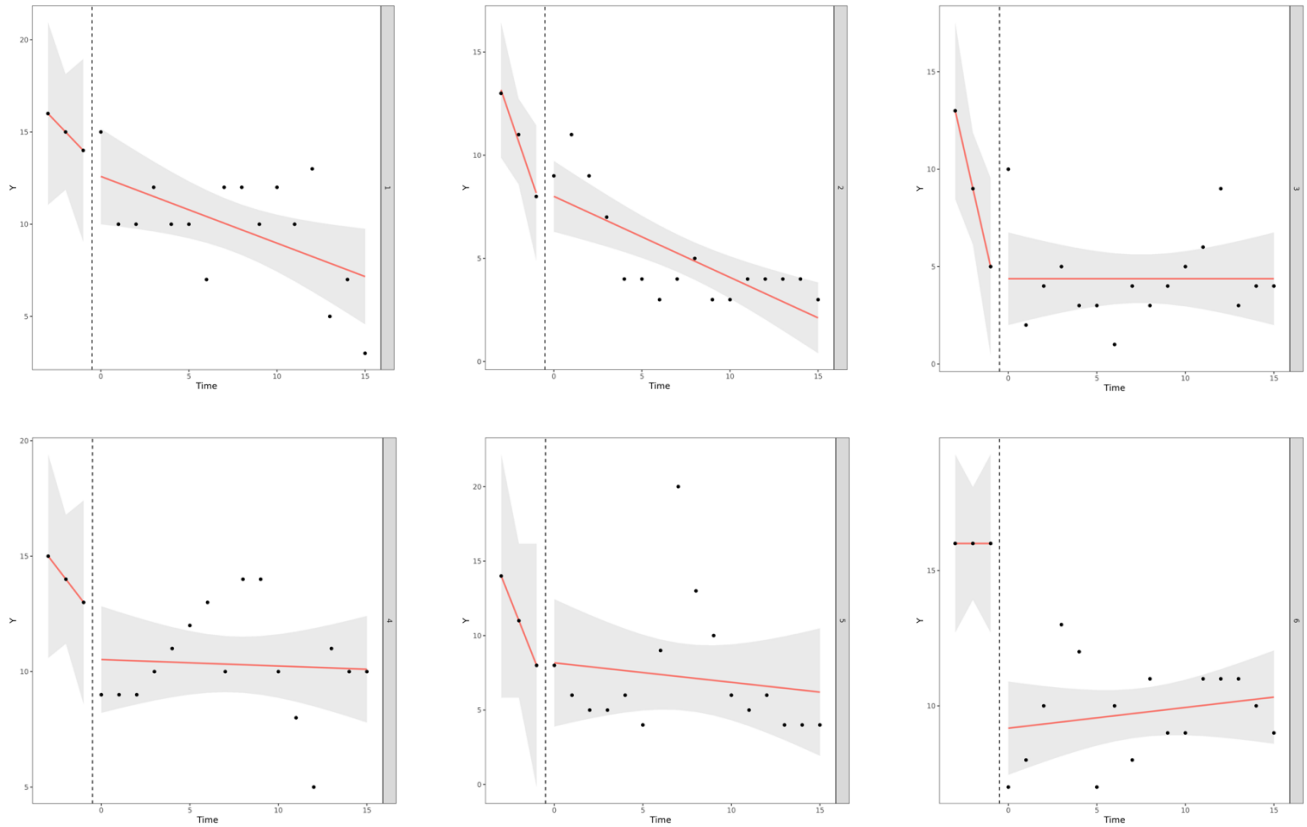
Concerning the change in mean from the baseline to the treatment phase, a difference can be observed in many cases such as: subject 1 (M_B= 15, $M_T = 9{,}88$), subject 2 (M_B= 10,7, $M_T = 5{,}06$), subject 3 (M_B= 9, $M_T = 4{,}4$), subject 5 (M_B= 11, $M_T = 7{,}19$), subject 6 (M_B= 16, $M_T = 9{,}75$), subject 7 (M_B= 13,7, $M_T = 10{,}43$), subject 12 (M_B= 7,7, $M_T = 1{,}4$).

Besides, it is possible to observe with the naked eye the shift in level from the last assessment of Phase 0 to the first assessment of Phase1 in subject 4, subject 6, subject 7, subject 10 and subject 12. These cases also present a negative score on coefficient of "Phase1" (Table 2) which indicates the immediate effect of treatment right after its start.

Contrarily, different considerations need to be done for the change in slope. In fact, in most of the cases (i.e., 1,2,3,4,5,8,9,12,13) the registered slope during the baseline phase (Time, $\hat{\beta}_1$) is steeper - in the desired direction - than the one in the treatment phase. This phenomenon interferes on the statistical analysis, since it seems that main effect on symptoms reduction is given by the baseline condition instead of the treatment one. On the other hand, it is essential to note that, also during the treatment condition, the symptoms scores keep decreasing or remain steady way under the cutoff of 10.

Similarly, the overall latency of change across cases is difficult to evaluate because the symptoms start decreasing during the baseline condition and keep decreasing during treatment.

Figure 3. *Graphical representation of the one-level model parameters from Eq.1, with confidence interval (α=0.05) and after centering the time variable.*



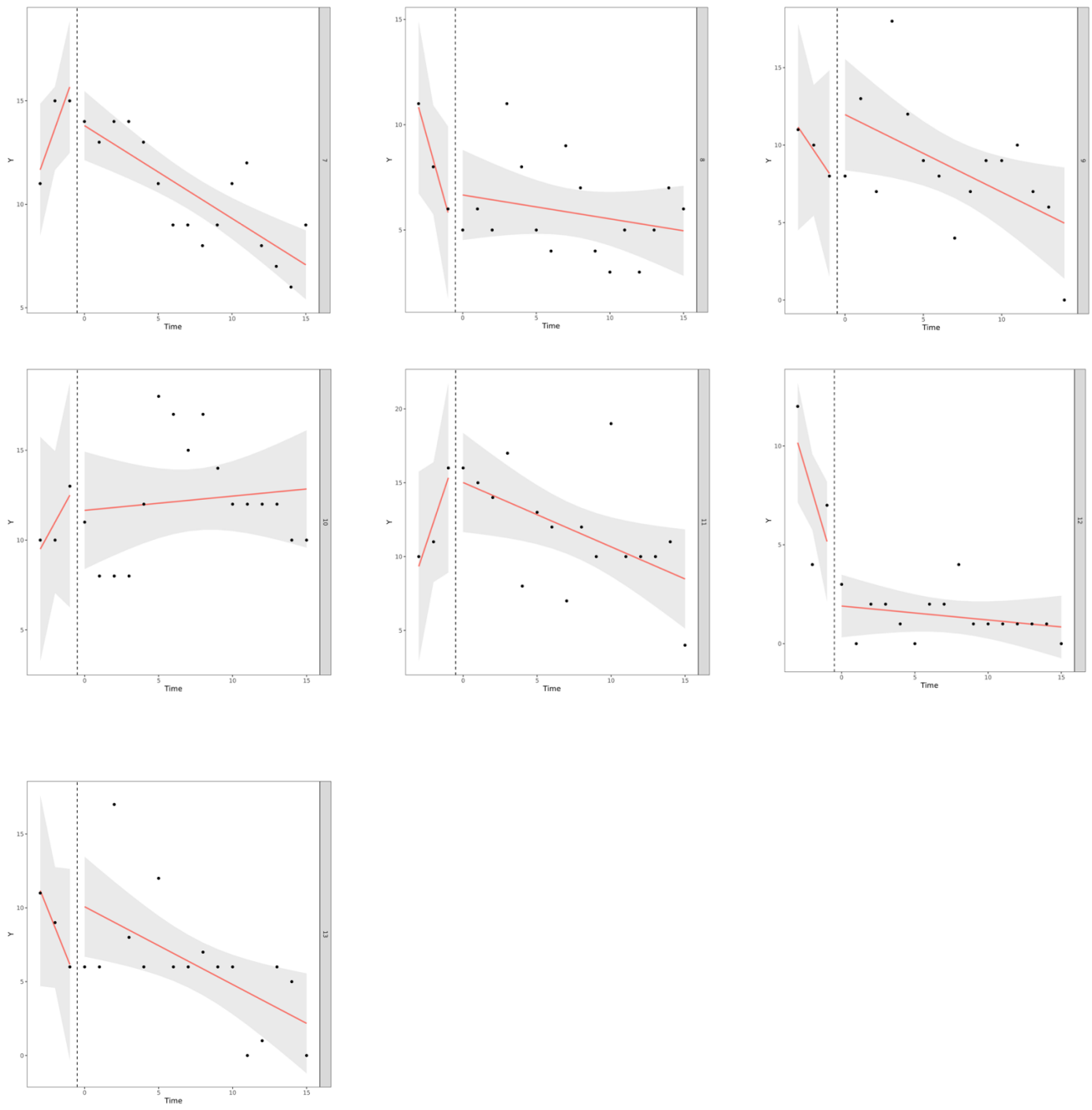


Figure 3.

4.2 Two-level analysis: intervention effect across cases

A total of 246 observations over 13 single cases were analyzed through a two-level analysis (Eq. 5). The numerical results are shown in Table 4 and presented graphically in Figure 4 (where the one-level and two-level regression analysis are compared). Across all the cases, the expected scores decrease on average by -1.077 per time unit if there is no intervention, down until 9.15 points [t (12) = 4.84, p <0.001] at the start of the treatment phase. The intervention has an average immediate effect of 0.372 points, and it increases the time trend by 0.859 points. However, neither the baseline trend (Time), the immediate intervention effect (Phase) and the effect of the treatment on the trend (Time:Phase1) are significant, meaning that it is not possible to conclude that they are different from zero.

Fixed effects

	Estimate	Std. Error	KR df	t-value	p-value
(Intercept)	9.1538462	1.8917610	12.00000	4.8387964	0.0004058799
Time	-1.0769231	0.7081706	12.00000	-1.5207113	0.1542355642
Phase1	0.3724648	1.3822323	11.99994	0.2694661	0.7921486147
Time:Phase1	0.8587850	0.7174615	11.99994	1.1969772	0.2544275698

Table 4. *Two-level regression results*

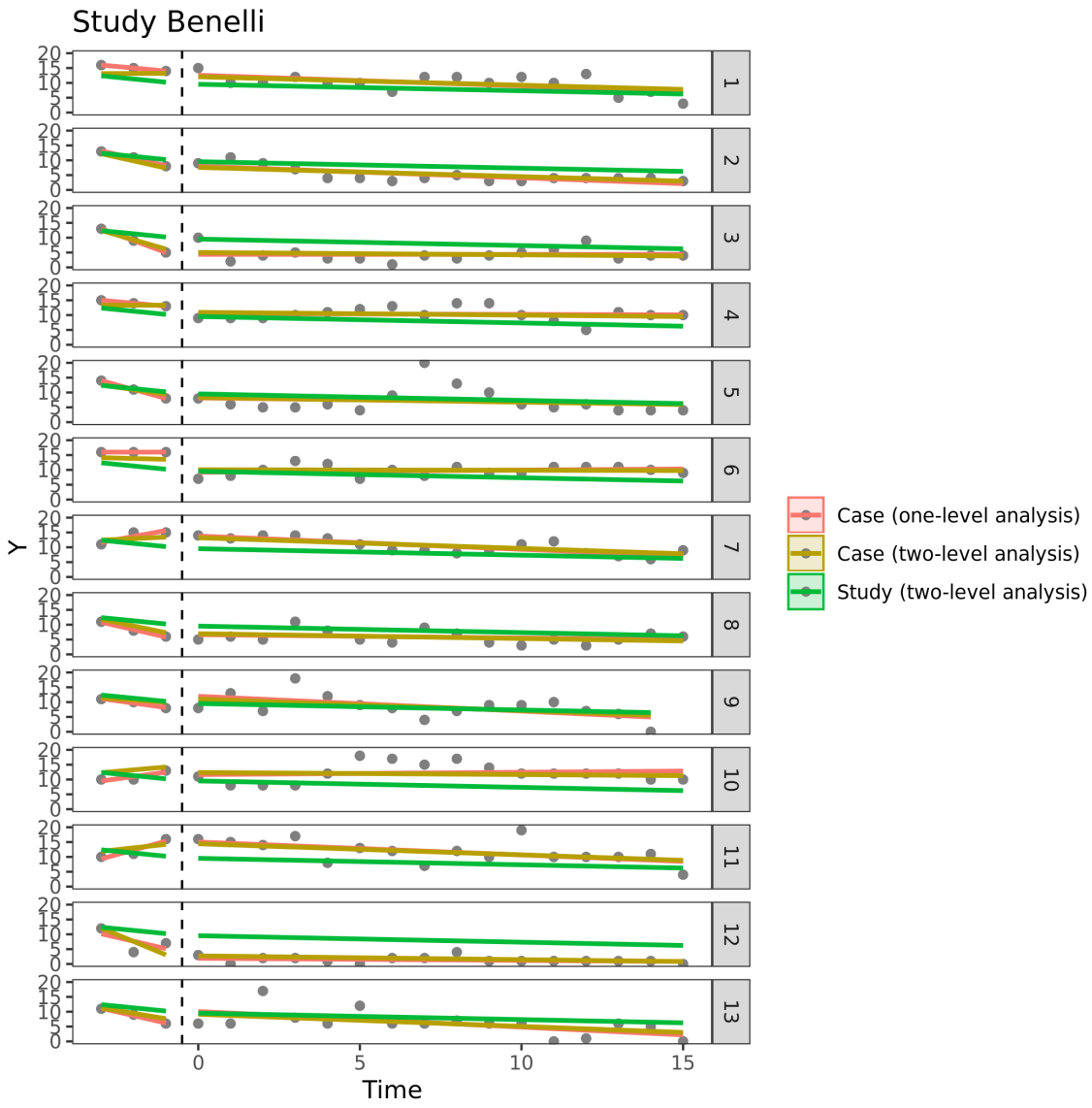


Figure 4. *Comparison of the one-level and two-level regression analysis.*

5. Discussion

In the present study, multilevel regression models were employed to meta-analyze a series of 13 single case experimental designs with the aim of supporting the efficacy of Transactional Analysis in the treatment of anxiety disorders. SCED data were analyzed *within cases* (one-level analysis) and *across cases* (two-level analysis). To our knowledge, this is the first research applying this methodology to examine the outcome of psychotherapy on the treatment of a common mental disorder in adults. For this purpose, the software MultiSCED, by Declercq et al. (2020), was adopted.

Results are mixed and they need to be interpreted cautiously. In the one-level analysis (Eq. 1) only 4/13 cases show significant regression coefficients. These are: subject 3,6,7,12. Specifically, subjects 3 and 12 show a significant decrease in the GAD7 scores already in the baseline phase which continues – although much less intensively – during the intervention phase. Undeniably, this creates some problems in the evaluation of the intervention consequences because it is in the same direction of the predicted treatment effects. Indeed, given that in the assessment phase the symptoms decrease (defined by the slope $\hat{\beta}_1$ of the Time variable) looks to be steeper, the influence of the psychotherapy treatment easily fades, even though the visual analysis clearly shows that the GAD7 scores keep decreasing or remain steady way under the clinical cutoff. The situation is different in subject 6 and 7, which present, respectively, a stable and an increasing baseline and two significant intercepts $\hat{\beta}_0$=16 and $\hat{\beta}_0$=17.667. These initial conditions foster a more straightforward interpretation of the treatment effects, that become directly responsible for the patient's change. In fact, subject 6 shows a significant coefficient of the regressor Phase1 ($\hat{\beta}_2 = -6.824, p < 0.05$), which indicates the immediate effect of the treatment; while, in subject 7, there is a significant decrease of symptoms on the time trend ($\hat{\beta}_3 = -2.449, p < 0.05$).

In addition to cases 3 and 12, also subjects 1,2,4,5,8,9,13 show an evident decrease on GAD7 scores already during the baseline phase (even though it is not significant in the one-level analysis). The fact that 9/13 patients present such an important change before the treatment is worth of note. As discussed also in paragraph 1.1.2, patients frequently exhibit noticeable changes already after the first

assessment sessions (Frank et al., 1963). These improvements, however, are due to a phenomenon called “restoration of hope” (e.g., “remoralization”) in the patient (Howard et al., 1993). Greenberg et al. (2006) provided an example of this, such as a therapist saying at the beginning of treatment: "It makes sense that you sought this type of help for your difficulties" or, "Depressions - in our case anxiety - do respond to treatment and the prognosis is quite good" (p. 671). On top of this, the intrinsic nature of anxiety disorders is probably prone to show immediate improvements during the initial stages of the intervention. Indeed, the simple fact of being listened by an emphatic and non-judgmental professional, together with setting therapy goals, gives back to patients an expectation to have “control” over their problems (in a life period characterized by always trying, in vane, to have everything under control) (Budge & Wampold, 2015).

Given the intrinsic complexity of the object of our research, results at the individual level were investigated considering also other perspectives. Indeed, it is important to remember that both clinicians and research have to do with living human beings whose inner worlds cannot be captured solely by a self-report. In this scenario, single case studies are getting more and more central since they allow to complement statistical considerations with other elements. For this reason, each patient was evaluated also through the reliable clinical change and the visual representation of the treatment progresses. The choice of adopting diverse approaches in the data analysis was driven by the awareness that there is a different between clinical and statistical significance. During a psychotherapy treatment even minor changes might have a big impact. Thus, more focus is essential to identify changes that are needed to produce a difference, not necessarily from a statistical point of view but, rather, from the perspective of the patient well-being in the real life (Kazdin, 2019).

To this end, visual inspection of the data helped us to observe closer all the changes within each patient. The one-level regression plots allowed to rapidly identify the change in mean between phases in the desired direction (happened in 7 cases), the immediate shift in level from the baseline to the intervention phase (happened in 5 cases) and the change in slope. Regarding this last point, the plots show that in 9/13 SCEDs the regression slope during the baseline phase is steeper than the one

in the treatment phase, indicating a higher decrease in symptoms before the start of the treatment. Despite this, in all these nine SCEDs, scores keep decreasing or remain steady under the clinical cutoff for GAD7 (10) in the intervention phase, meaning a lighter, but still present, treatment effect.

According to the developers of the *reliable change index* (RCI) Jacobson & Truax (1991), whether a treatment effect exists in the statistical sense has little to do with the clinical significance of the effect. This is very true for the current study. Although only a few SCEDs in this sample show significant results in the one-level analysis, 9/13 patients experienced clinical reliable change, with a decrease of six or more points from the first baseline assessment to the last treatment session (Bischoff et al., 2020).

Finally, the effect of the intervention across cases was calculated through a second-level analysis (Eq. 5). Unfortunately, none of the estimates reached the statistical significance. Thus, based on the current findings, we cannot affirm that the changes happened in our sample can be attributed to the intervention effects. In other words, even though the majority of SCEDs showed a reliable clinical change at the individual level, it is not possible to draw any conclusion regarding the efficacy of Transactional Analysis in treating anxiety disorders and further research is needed.

6. Conclusions

6.1 Limitations

The current study presents some limitations that had a big impact on the final outcome, particularly on the regression analysis. Before describing them in detail, it is important to specify that the analysis was not pre-registered on Open Science Framework (OSF) and the PRISMA guidelines were not followed.

Firstly, the number of assessments in the baseline phase was not sufficient to obtain reliable results both in the one-level and two-level analysis. If it is true that a baseline is conventionally defined as a minimum of three-data-points recorded before the beginning of a treatment (Kazdin, 2010), it is also noteworthy to fully consider the method employed in our analysis. Indeed, three measurements in the assessment condition cannot guarantee - in most cases - the stability of the baseline. This is rather an important requirement in the evaluation of the treatment effects through multilevel modeling, which is fundamentally based on regression equations. Given that baseline performance is used to predict how the client will respond in the immediate future without intervention, it is crucial that the data are relatively stable. Thus, the desired assessment scores should display an absence of a trend or little fluctuation (Kazdin, 2019). In our sample, this feature is present only in subject 6 which, unsurprisingly, shows a net decrease on the GAD7 scores right after the start of the treatment phase. In all the other cases (except for subjects 7, 10, 11) there is a trend towards the decrease of symptoms already during baseline phase; this creates some evaluation issues because it is in the same direction of the predicted treatment effects (Kazdin, 2019). Contrarily, case 7 and 11, which present an increase on the GAD7 scores in the baseline phase, show a clear decrease in the treatment phase, given by the intervention effects on the time trend (in subject 7, the Time:Phase1 regression coefficient is even significant).

The same reasoning about stability can be applied also to the follow-up. In fact, a considerable number of cases in our sample also reported three assessments of the follow up phase. However,

given that three measurements can't predict a stable performance, we decided not to include them in the analysis (an example of how MultiSCED treats follow-up data can be found in the Appendix E).

The second important issue of the current study regards the measurement instrument. Although the Generalized Anxiety Disorder-7 scale (Spitzer et al., 2006) is widely used both in clinical practice and for research purposes, it might not be sensitive enough to detect small intrapersonal changes that are typical of continuous assessment; indeed, GAD7 present an error on raw data of 6/21 (around 28,6%).

Furthermore, a lot of SCEDs in the current research show several outlier points that negatively influenced the result. In fact, it has been attempted to delete 1 outlier from subject 3 and subject 11; in both cases, the elimination of only one divergent data point led to obtain significant regression coefficients on Phase and Time:Phase1.

Finally, another key issue of the current study regards the size of our sample. Moeyaert et al. (2020) recommended being cautious when fewer than 30 single cases are available, given that such situations are related to less accurate fixed effects estimates and less power. Besides, the authors suggested selecting cases with around 20 measurements per individual (Moeyaert et al., 2020). The present study started with an initial dataset of 41 SCEDs. However, many of them had an insufficient number of measurements in both phases or showed a subclinical score in the baseline phase. Thus, we decided to prefer quality over quantity and selecting only cases that had enough measurements to make reliable predictions (all the SCEDs in this work have 19 measurements, except for case 9 which has 18). In doing so, the general size of the sample was sacrificed, resulting in non-significant estimates in the two-level analysis.

6.2 Conclusions and Future Directions

This study was conducted with the three objectives: a) evaluating the efficacy of Transactional Analysis in treating anxiety disorders, b) shading light on the application of multilevel modeling for the meta-analysis of single case experimental designs (SCEDs), c) testing the MultiSCED software (Declercq et al., 2020) as an innovative instrument to bridge the gap between research and clinical practice.

Unfortunately, the first goal was not met. Despite a consistent number of SCEDs in our sample showed a reduction in symptoms and a reliable clinical change, it is not possible to draw any final consideration regarding the efficacy of Transactional Analysis in treating anxiety disorder. Given the limitations of the present data (e.g., too few baseline measurements, small sample size, etc., see paragraph 6.1), other previously tested statical methods (cf. Benelli & Zanchetta, 2019) could be employed to compute the SCEDs meta-analysis, such as the standardized mean difference statistic (*d*) (Shadish et al., 2014).

Regarding the second objective, multilevel modeling demonstrated to be specifically suitable for summarizing hierarchical structured data, such as single case experimental designs (SCEDs) (Van Den Noortgate & Onghena, 2003). Its strength stands in the possibility to summarize participant-specific effect sizes across cases and across studies (Moeyaert, 2019). However, as the present study illustrates, some precautions must be taken before embarking on this method. Firstly, at least 30 SCEDs must be available, each one needs to have around 20 measurements (Moeyaert et al. 2020). Secondly, enough measurements must be recorded during the initial assessment to guarantee a stable baseline.

In the context of psychotherapy outcome research, one could wonder whether an assessment phase lasting more than three sessions (without the delivery of treatment), could be considered ethical and beneficial for the final outcome. A possible solution to this matter would be changing the assessment strategy with outpatient population. Nowadays, technology (e.g., use of apps and smartphones) comes in handy, potentially making daily assessment feasible for everyone.

For instance, a recent SCED study by Bottesi et al. (2023) employed an ad-hoc daily self-monitoring questionnaire for measuring the primary outcome. On the contrary, multilevel modeling methodology is more easily applicable in inpatients settings (e.g., hospitals), where continuous assessment usually means daily assessment (Kazdin, 2019).

As already mentioned in the previous paragraph, in the case of continuous assessment, the measurement instrument has to be adjusted to make sure it will be sensitive enough to detect even small changes.

A higher number of measurements would also leave room for correcting for outliers. Indeed, as reported in paragraph 6.1, the elimination of only one divergent data point led to obtain significant regression coefficients in two SCEDs of the current study.

Since the present work was intended to be exploratory, further research is needed on this topic, taking into consideration all the above-mentioned issues and proposed solutions.

Finally, the MultiSCED software confirmed itself to be a cutting-edge instrument that could certainly help to bridge the gap between research and clinical practice. Thanks to its clear interface, clinicians and researchers are made able to test and summarize the effects of their work. Indeed, the utility of this tool ranges from the single psychotherapist wanting to show patients their progresses, until a research team aiming to prove the efficacy of an emerging treatment through a meta-analysis.

7. References

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*(4), 271–285. https://doi.org/10.1037/0003-066X.61.4.271

Baek, E., Luo, W., & Lam, K. H. (2023). Meta-Analysis of Single-Case Experimental Design using Multilevel Modeling. *Behavior Modification*, 014544552211440. https://doi.org/10.1177/01454455221144034

Benelli, E., Paolillo, A. R., & Ventriglia, S. (2021). *Analisi transazionale per i disturbi ansiosi. Manuale per il trattamento*. FrancoAngeli.

Benelli, E., & Zanchetta, M. (2019). Single-Case Design Review and Meta-Analysis for Supporting the Method of Transactional Analysis towards Recognition as an Empirically Supported Treatment for Depression. *International Journal of Psychotherapy*, *23*(3), 93–108.

Bergmann, L. H. (1981). A Cognitive Behavioral Approach to Transactional Analysis. *Transactional Analysis Journal*, *11*(2), 147–149. https://doi.org/10.1177/036215378101100211

Berne, E. (1958). Transactional Analysis: A New and Effective Method of Group Therapy. *American Journal of Psychotherapy*, *12*(4), 735–743. https://doi.org/10.1176/appi.psychotherapy.1958.12.4.735

Berne, E. (1961). *Transactional analysis in psychotherapy: A systematic individual and social psychiatry.* Grove Press. https://doi.org/10.1037/11495-000

Bischoff, T., Anderson, S. R., Heafner, J., & Tambling, R. (2020). Establishment of a reliable change index for the GAD-7. *Psychology, Community & Health*, *8*(1), 176–187. https://doi.org/10.5964/pch.v8i1.309

Bottesi, G., Contin, S. A., Panzeri, A., Carraro, E., Bianconi, S., & Ghisi, M. (2023). Un intervento transdiagnostico di gruppo focalizzato sull'intolleranza dell'incertezza. *Italian Journal of Cognitive and Behavioural Psychotherapy*, *29*(1), 47–69.

Bower, P., & Gilbody, S. (2005). Stepped care in psychological therapies: access, effectiveness and efficiency. *British Journal of Psychiatry*, *186*(1), 11–17. https://doi.org/10.1192/bjp.186.1.11

Braakmann, D. (2015). Historical Paths in Psychotherapy Research. In O. C. G. Gelo, B. Rieken, & A. Pritz (Eds.), *Psychotherapy Research. Foundations, Process, and Outcome* (pp. 39–66). Springer.

Budge, S. L., & Wampold, B. E. (2015). The Relationship: How It Works. In *Psychotherapy Research* (pp. 213–228). Springer Vienna. https://doi.org/10.1007/978-3-7091-1382-0_11

Cawthorne, T., Käll, A., Bennett, S., Baker, E., Cheung, E., & Shafran, R. (2023). Do single-case experimental designs lead to randomised controlled trials of cognitive behavioural therapy interventions for adolescent anxiety and related disorders recommended in the National Institute of Clinical Excellence guidelines? A systematic review. *JCPP Advances*. https://doi.org/10.1002/jcv2.12181

Chambless, D. L., & Hollon, S. D. (1998). Defining Empirically Supported Therapies. In *Journal of Consulting and Clinical Psychology* (Vol. 66, Issue 1).

Chambless, D. L., & Ollendick, T. H. (2001). Empirically Supported Psychological Interventions: Controversies and Evidence. *Annual Review of Psychology*, *52*(1), 685–716. https://doi.org/10.1146/annurev.psych.52.1.685

Declercq, L., Cools, W., Beretvas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (2020). MultiSCED: A tool for (meta-)analyzing single-case experimental data with multilevel modeling. *Behavior Research Methods*, *52*(1), 177–192. https://doi.org/10.3758/s13428-019-01216-2

DeFife, J., Drill, R., Beinashowitz, J., Ballantyne, L., Plant, D., Smith-Hansen, L., Teran, V., Werner-Larsen, L., Westerling, T., Yang, Y., Davila, M., & Nakash, O. (2015). Practice-based psychotherapy research in a public health setting: Obstacles and opportunities. *Journal of Psychotherapy Integration*, *25*(4), 299–312. https://doi.org/10.1037/a0039564

Delgadillo, J., Overend, K., Lucock, M., Groom, M., Kirby, N., McMillan, D., Gilbody, S., Lutz, W., Rubel, J. A., & de Jong, K. (2017). Improving the efficiency of psychological treatment using outcome feedback technology. *Behaviour Research and Therapy*, *99*, 89–97. https://doi.org/10.1016/j.brat.2017.09.011

Elliott, R. (2002). Hermeneutic single-case efficacy design. *Psychotherapy Research*, *12*(1), 1–21. https://doi.org/10.1080/713869614

EVANS, J. G. (1995). Evidence-based and Evidence-biased Medicine. *Age and Ageing*, *24*(6), 461–463. https://doi.org/10.1093/ageing/24.6.461

Evidence-based practice in psychology. (2006). *American Psychologist*, *61*(4), 271–285. https://doi.org/10.1037/0003-066X.61.4.271

Firth, N., Barkham, M., & Kellett, S. (2015). The clinical effectiveness of stepped care systems for depression in working age adults: A systematic review. *Journal of Affective Disorders*, *170*, 119–130. https://doi.org/10.1016/j.jad.2014.08.030

FRANK, J. D., NASH, E. H., STONE, A. R., & IMBER, S. D. (1963). IMMEDIATE AND LONG-TERM SYMPTOMATIC COURSE OF PSYCHIATRIC OUTPATIENTS. *American Journal of Psychiatry*, *120*(5), 429–439. https://doi.org/10.1176/ajp.120.5.429

Gaynor, S. T., & Harris, A. (2008). Single-Participant Assessment of Treatment Mediators. *Behavior Modification*, *32*(3), 372–402. https://doi.org/10.1177/0145445507309028

Gold, C. (2015). Quantitative Psychotherapy Outcome Research: Methodological Issues. In O. C. G. Gelo, B. Rieken, & Pritz Alfred (Eds.), *Psychotherapy Research. Foundations, Process, and Outcome* (pp. 537–558). Springer.

Goodheart, C. D., Levant, R. F., Barlow, D. H., Carter, J., Davidson, K. W., Hagglund, K. J., Hollon, S. D., Johnson, J. D., Leviton, L. C., Mahrer, A. R., Newman, F. L., Norcross, J. C., Silverman, D. K., Smedley, B. D., Wampold, B. E., Westen, D. I., Yates, B. T., Zane, N. W., Reed, G. M., … Bullock, M. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*(4), 271–285. https://doi.org/10.1037/0003-066X.61.4.271

Greenberg, R. P., Constantino, M. J., & Bruce, N. (2006). Are patient expectations still relevant for psychotherapy process and outcome? *Clinical Psychology Review*, *26*(6), 657–678. https://doi.org/10.1016/j.cpr.2005.03.002

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Hargaden, H., & Sills, C. (2014). *Transactional Analysis*. Routledge. https://doi.org/10.4324/9781315820279

Hornstra, R., Onghena, P., van den Hoofdakker, B. J., van der Veen-Mulders, L., Luman, M., Staff, A. I., & van der Oord, S. (2023). Components of Behavioral Parent Training for Children With Attention-Deficit/Hyperactivity Disorder: A Series of Replicated Single-Case Experiments. *Behavior Modification*. https://doi.org/10.1177/01454455231162003

Howard, K. I., Lueger, R. J., Maling, M. S., & Martinovich, Z. (1993). A phase model of psychotherapy outcome: Causal mediation of change. *Journal of Consulting and Clinical Psychology*, *61*(4), 678–685. https://doi.org/10.1037/0022-006X.61.4.678

Ilardi, S. S., & Craighead, W. E. (1994). The role of nonspecific factors in cognitive-behavior therapy for depression. *Clinical Psychology: Science and Practice*, *1*(2), 138–156. https://doi.org/10.1111/j.1468-2850.1994.tb00016.x

Jacobson, N. S., & Truax, P. (1991). *Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research* (Vol. 59, Issue 1).

Kazdin, A. E. (1981). Drawing valid inferences from case studies. *Journal of Consulting and Clinical Psychology*, *49*(2), 183–192. https://doi.org/10.1037/0022-006X.49.2.183

Kazdin, A. E. (2010). *KAZDIN, A.E. (2010). Single-case research designs: Methods for clinical and applied settings (2nd ed.). New York. Oxford University Press* (2nd ed.). Oxford University Press.

Kazdin, A. E. (2011). *Kazdin, A. E. (2011). Single-Case Research Designs* (2nd ed.). Oxford University Press. https://doi.org/10.1080/07317107.2012.654458

Kazdin, A. E. (2019). Single-case experimental designs. Evaluating interventions in research and clinical practice. *Behaviour Research and Therapy*, *117*, 3–17. https://doi.org/10.1016/j.brat.2018.11.015

Kazdin, A. E. (2021). Single-case experimental designs: Characteristics, changes, and challenges. *Journal of the Experimental Analysis of Behavior*, *115*(1), 56–85. https://doi.org/10.1002/jeab.638

Kocsis, J. H., Leon, A. C., Markowitz, J. C., Manber, R., Arnow, B., Klein, D. N., & Thase, M. E. (2009). Patient Preference as a Moderator of Outcome for Chronic Forms of Major Depressive Disorder Treated With Nefazodone, Cognitive Behavioral Analysis System of Psychotherapy, or Their Combination. *The Journal of Clinical Psychiatry*, *70*(3), 354–361. https://doi.org/10.4088/JCP.08m04371

Kratochwill, T. R., & Levin, J. R. (2014). *Single-case intervention research: Methodological and statistical advances.* American Psychological Association. https://doi.org/10.1037/14376-000

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., & Löwe, B. (2007). Anxiety Disorders in Primary Care: Prevalence, Impairment, Comorbidity, and Detection. *Annals of Internal Medicine*, *146*(5), 317. https://doi.org/10.7326/0003-4819-146-5-200703060-00004

Lambert, M., & Ogles, B. (2004). The efficacy and effectiveness of psychotherapy. In Wiley (Ed.), *Lambert M (ed) Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 139–193).

Maric, M., Schumacher, L., Van den Noortgate, W., Bettelli, L., Engelbertink, W., & Stikkelbroek, Y. (2023). A Multilevel Meta-analysis of Single-Case Research on Interventions for Internalizing Disorders in Children and Adolescents. *Clinical Child and Family Psychology Review*, *26*(2), 416–429. https://doi.org/10.1007/s10567-023-00432-9

Maric, M., Wiers, R. W., & Prins, P. J. M. (2012). Ten Ways to Improve the Use of Statistical Mediation Analysis in the Practice of Child and Adolescent Treatment Research. *Clinical Child and Family Psychology Review*, *15*(3), 177–191. https://doi.org/10.1007/s10567-012-0114-y

Mattias Desmet. (2013). Experimental versus naturalistic psychotherapy research: consequences for researchers, clinicians, policy makers and patients. *Psychoanalytische Perspectieven* , *31*(1), 59–78.

Meganck, R., Krivzov, J., Notaerts, L., Willemsen, J., Kaluzeviciute, G., Dewaele, A., & Desmet, M. (2022). The single case archive: Review of a multitheoretical online database of published peer-reviewed single-case studies. *Psychotherapy*, *59*(4), 641–646. https://doi.org/10.1037/pst0000431

Moeyaert, M. (2019). Quantitative Synthesis of Research Evidence: Multilevel Meta-Analysis. *Behavioral Disorders*, *44*(4), 241–256. https://doi.org/10.1177/0198742918806926

Moeyaert, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology*, *52*(2), 191–211. https://doi.org/10.1016/j.jsp.2013.11.003

Moeyaert, M., Manolov, R., & Rodabaugh, E. (2020). Meta-Analysis of Single-Case Research via Multilevel Models: Fundamental Concepts and Methodological Considerations. In *Behavior Modification* (Vol. 44, Issue 2, pp. 265–295). SAGE Publications Inc. https://doi.org/10.1177/0145445518806867

Moiso, C., & Novellino, M. (2000). An Overview of the Psychodynamic School of Transactional Analysis and Its Epistemological Foundations. *Transactional Analysis Journal*, *30*(3), 182–187. https://doi.org/10.1177/036215370003000302

Nikolakopoulou, A., Mavridis, D., & Salanti, G. (2014). How to interpret meta-analysis models: Fixed effect and random effects meta-analyses. *Evidence-Based Mental Health*, *17*(2), 64. https://doi.org/10.1136/eb-2014-101794

Philips, B., & Falkenström, F. (2021). What Research Evidence Is Valid for Psychotherapy Research? *Frontiers in Psychiatry*, *11*. https://doi.org/10.3389/fpsyt.2020.625380

Population Health, Clinical Audit, Specialist Care Team, & NHS Digital. (2022, September 29). *Psychological Therapies, Annual report on the use of IAPT services, 2021-22*. Https://Digital.Nhs.Uk/Data-and-Information/Publications/Statistical/Psychological-Therapies-Annual-Reports-on-the-Use-of-Iapt-Services/Annual-Report-2021-22.

Raue, P. J., Schulberg, H. C., Heo, M., Klimstra, S., & Bruce, M. L. (2009). Patients' Depression Treatment Preferences and Initiation, Adherence, and Outcome: A Randomized Primary Care Study. *Psychiatric Services*, *60*(3), 337–343. https://doi.org/10.1176/ps.2009.60.3.337

Reese, R. J., Norsworthy, L. A., & Rowlands, S. R. (2009). Does a continuous feedback system improve psychotherapy outcome? *Psychotherapy*, *46*(4), 418–431. https://doi.org/10.1037/a0017901

Rindskopf, D. M., & Ferron, J. M. (n.d.). Using multilevel models to analyze single-case design data. In D. M. Rindskopf & J. M. Ferron (Eds.), *Single-case intervention research: Methodological and statistical advances.* (pp. 221–246). American Psychological Association. https://doi.org/10.1037/14376-008

Schuurman, N. K. (2023). *A " Within/Between Problem" Primer: About (not) separating within-person variance and between-person variance in psychology*.

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology*, *52*(2), 109–122. https://doi.org/10.1016/j.jsp.2013.11.009

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology*, *52*(2), 123–147. https://doi.org/10.1016/j.jsp.2013.11.005

Shean, G. D. (2012). Some Limitations on the External Validity of Psychotherapy Efficacy Studies and Suggestions for Future Research. *American Journal of Psychotherapy*, *66*(3), 227–242. https://doi.org/10.1176/appi.psychotherapy.2012.66.3.227

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*(4), 510–550. https://doi.org/10.1037/a0029312

Solgi, Z., Falah Nodehi, M., Khalili, N., Mousavi, S., & Solgi, 1 Zahra. (2021). The effectiveness of transactional analysis psychotherapy on negative automatic thoughts and optimism of female adolescents with social anxiety disorder. *Journal of Research in Psychopathology*, *2*(6). https://doi.org/10.22098/jrp.2022.10092.1047

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder. *Archives of Internal Medicine*, *166*(10), 1092. https://doi.org/10.1001/archinte.166.10.1092

Stiles, W. B., Hill, C. E., & Elliott, R. (2015). Looking both ways. *Psychotherapy Research*, *25*(3), 282–293. https://doi.org/10.1080/10503307.2014.981681

Summers, G., & Tudor, K. (2018). *Co-creative transactional analysis: Papers, responses, dialogues, and developments.* Routledge.

Tolin, D. F., Mckay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically Supported Treatment: Recommendations for a New Model. In *Clinical Psychology: Science and Practice* (Vol. 22, Issue 4, pp. 317–338). Blackwell Publishing Inc. https://doi.org/10.1111/cpsp.12122

Truijens, F. L. (2017). Do the numbers speak for themselves? A critical analysis of procedural objectivity in psychotherapeutic efficacy research. *Synthese*, *194*(12), 4721–4740. https://doi.org/10.1007/s11229-016-1188-8

Van Den Noortgate, W., & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 1–10. https://doi.org/10.3758/BF03195492

Vlaeyen, J. W. S., Wicksell, R. K., Simons, L. E., Gentili, C., De, T. K., Tate, R. L., Vohra, S., Punja, S., Linton, S. J., Sniehotta, F. F., & Onghena, P. (2020). From Boulder to Stockholm in

70 Years: Single Case Experimental Designs in Clinical Research. *Psychological Record*, *70*(4), 659–670. https://doi.org/10.1007/s40732-020-00402-5

Vos, J., & Van Rijn, B. (2021). The Evidence-Based Conceptual Model of Transactional Analysis: A Focused Review of the Research Literature. *Transactional Analysis Journal*, *51*(2), 160–201. https://doi.org/10.1080/03621537.2021.1904364

Vos, J., & Van Rijn, B. (2021b). The Transactional Analysis Review Survey: An Investigation Into Self-Reported Practices and Philosophies of Psychotherapists. *Transactional Analysis Journal*, *51*(2), 111–126. https://doi.org/10.1080/03621537.2021.1904355

Vos, J., & van Rijn, B. (2022). The Effectiveness of Transactional Analysis Treatments and Their Predictors: A Systematic Literature Review and Explorative Meta-Analysis. *Journal of Humanistic Psychology*. https://doi.org/10.1177/00221678221117111

Wakefield, S., Kellett, S., Simmonds-Buckley, M., Stockton, D., Bradbury, A., & Delgadillo, J. (2021). Improving Access to Psychological Therapies (IAPT) in the United Kingdom: A systematic review and meta-analysis of 10-years of practice-based evidence. *British Journal of Clinical Psychology*, *60*(1), 1–37. https://doi.org/10.1111/bjc.12259

Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The Empirical Status of Empirically Supported Psychotherapies: Assumptions, Findings, and Reporting in Controlled Clinical Trials. *Psychological Bulletin*, *130*(4), 631–663. https://doi.org/10.1037/0033-2909.130.4.631

Wolfe, K., Barton, E. E., & Meadan, H. (2019). Systematic Protocols for the Visual Analysis of Single-Case Research Data. *Behavior Analysis in Practice*, *12*(2), 491–502. https://doi.org/10.1007/s40617-019-00336-7

Appendix

A. Formula expressions used in the R implementation

Declercq et al. (2020) employed the present syntax to obtain respectively the one-level, two-level and three level regression models in the MultiSCED.

One-level analysis
Y ~ 1 + Time + Phase + Phase:Time

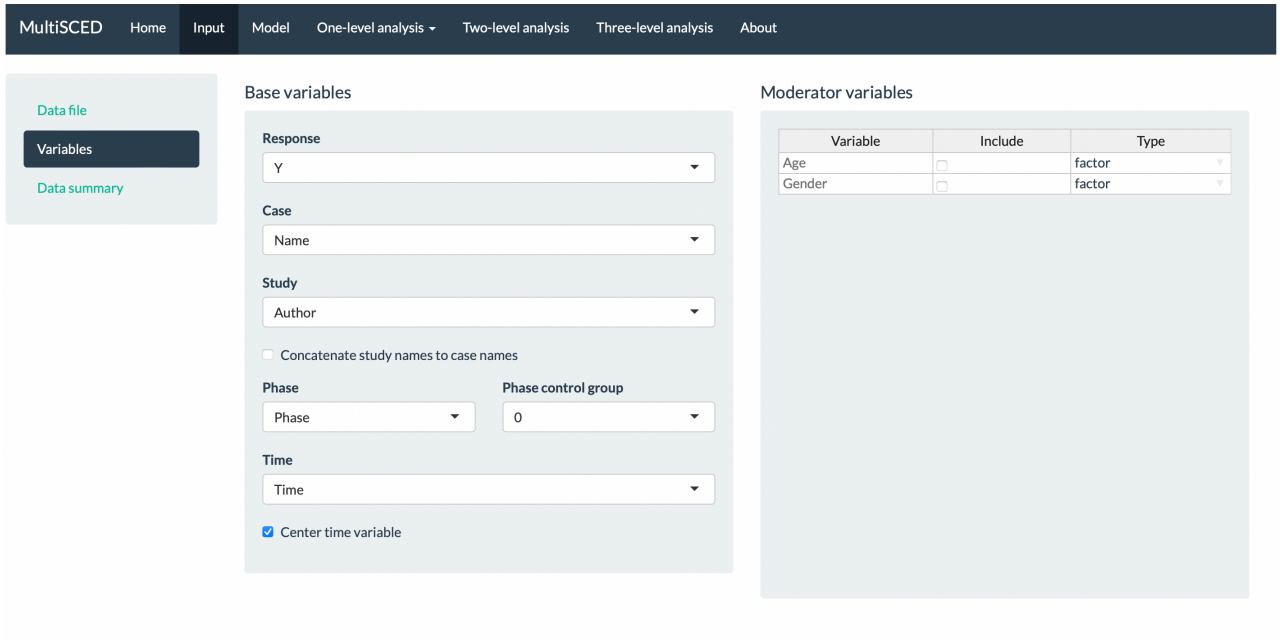
Two-level analysis
Y ~ 1 + Time + Phase + Phase:Time + (1 + Time + Phase + Phase:Time | Name)

Three-level analysis
Y ~ 1 + Time + Phase + Phase:Time + (1 + Time + Phase + Phase:Time | Author) + (1 + Time + Phase + Phase:Time | Author:Name)

B. Input data file

Author	Name	Age	Gender	Time	Phase	Y
Benelli	1	18	1	1	0	16
Benelli	1	18	1	2	0	15
Benelli	1	18	1	3	0	14
Benelli	1	18	1	4	1	15
Benelli	1	18	1	5	1	10
Benelli	1	18	1	6	1	10
Benelli	1	18	1	7	1	12
Benelli	1	18	1	8	1	10
Benelli	1	18	1	9	1	10
Benelli	1	18	1	10	1	7
Benelli	1	18	1	11	1	12
Benelli	1	18	1	12	1	12
Benelli	1	18	1	13	1	10
Benelli	1	18	1	14	1	12
Benelli	1	18	1	15	1	10
Benelli	1	18	1	16	1	13
Benelli	1	18	1	17	1	5
Benelli	1	18	1	18	1	7
Benelli	1	18	1	19	1	3
Benelli	2	18	1	1	0	13
Benelli	2	18	1	2	0	11
Benelli	2	18	1	3	0	8
Benelli	2	18	1	4	1	9
Benelli	2	18	1	5	1	11
Benelli	2	18	1	6	1	9
Benelli	2	18	1	7	1	7
Benelli	2	18	1	8	1	4
Benelli	2	18	1	9	1	4
Benelli	2	18	1	10	1	3
Benelli	2	18	1	11	1	4
Benelli	2	18	1	12	1	5
Benelli	2	18	1	13	1	3
Benelli	2	18	1	14	1	3
Benelli	2	18	1	15	1	4
Benelli	2	18	1	16	1	4
Benelli	2	18	1	17	1	4
Benelli	2	18	1	18	1	4
Benelli	2	18	1	19	1	3

Example of the input data file. The meta-analytic data were stored in a tab-delimited text (.txt) file with each row representing an observation and each column representing a variable.



Variable selection interface from the “Input” page in the MultiSCED (Declercq et al., 2020).

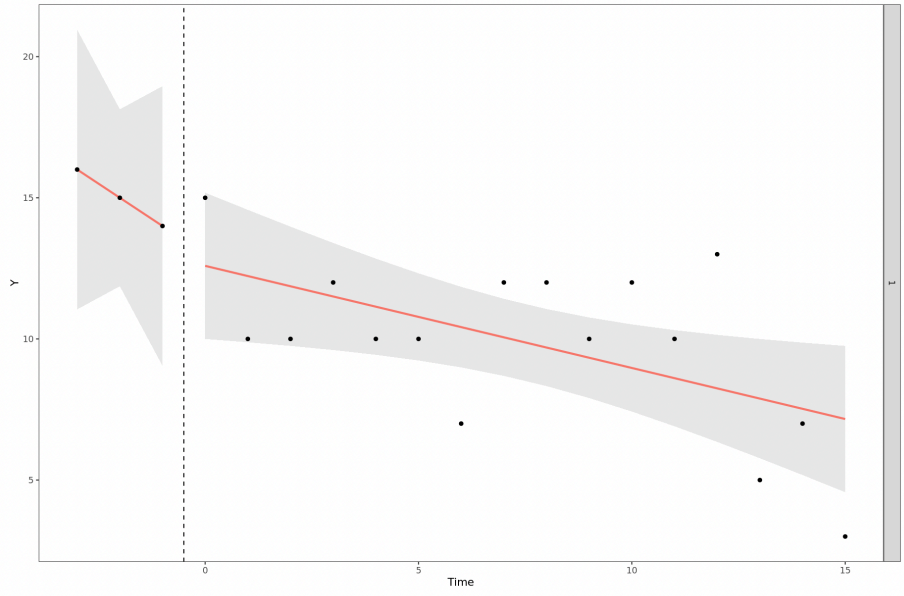
C. Comment of the one-level analysis output for each participant

Subject 1

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	1	(Intercept)	13	3.89	3.342	0.004
2	1	Time	-1	1.8	-0.555	0.587
3	1	Phase1	-0.412	4.075	-0.101	0.921
4	1	Time:Phase1	0.638	1.806	0.353	0.729

This patient, female, 18 years old, gets a diagnosis of a *social anxiety disorder* during the phase of assessment together with avoidant traits. According to Bischoff et al. (2020)'s criteria for determining GAD7 reliable change, it is possible to affirm that the patient experienced clinically significant change given that the first registered score on GAD7 during the assessment phase was 16 and the last registered score during the treatment phase was 3.

After conducting the one-level analysis it can be noticed that the estimated intercept β_0= 13 is statistically significant. This implies that if the baseline continued, the patient would have had a score of 13 at the start of the treatment. Unfortunately, all the other regression coefficients are not significant, hence it is impossible to affirm that the treatment effect are statistically different from zero. This could be due to the high number of outliers.



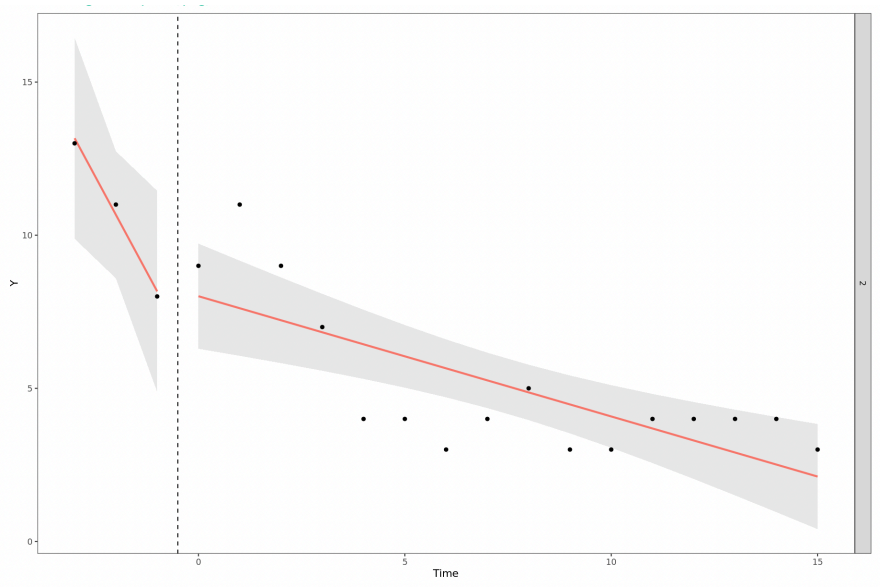
Subject	Comment
2	

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	2	(Intercept)	5.667	2.577	2.199	0.044
2	2	Time	-2.5	1.193	-2.096	0.053
3	2	Phase1	2.341	2.7	0.867	0.4
4	2	Time:Phase1	2.107	1.196	1.762	0.099

This patient, female, 18 years old, gets a diagnosis of a *panic attacks* during the phase of assessment. According to Bischoff et al. (2020) 's criteria for determining GAD7 reliable change, it is possible to affirm that the patient experienced clinically significant change given that the first registered score on GAD7 during the assessment phase was 13 and the last registered score during the treatment phase was 3.

After conducting the one-level analysis it can be noticed that the estimated intercept β_0= 5.667 is statistically significant. This implies that if the baseline continued, the patient would have had a score of 5.667 on GAD7 at the start of the intervention phase and the latter, without any treatment, would potentially decrease by - 2.5 points (β_1 = -2.5, p =0.053), with each additional time unit.

On the contrary, both coefficients of Phase1 and Time:Phase1 don't reach the statistical significance. This means that the start of the intervention doesn't have an immediate impact on the symptoms scores and that the treatment effect on the time trend is not statistically different from zero. This could be due to the high number of outliers.

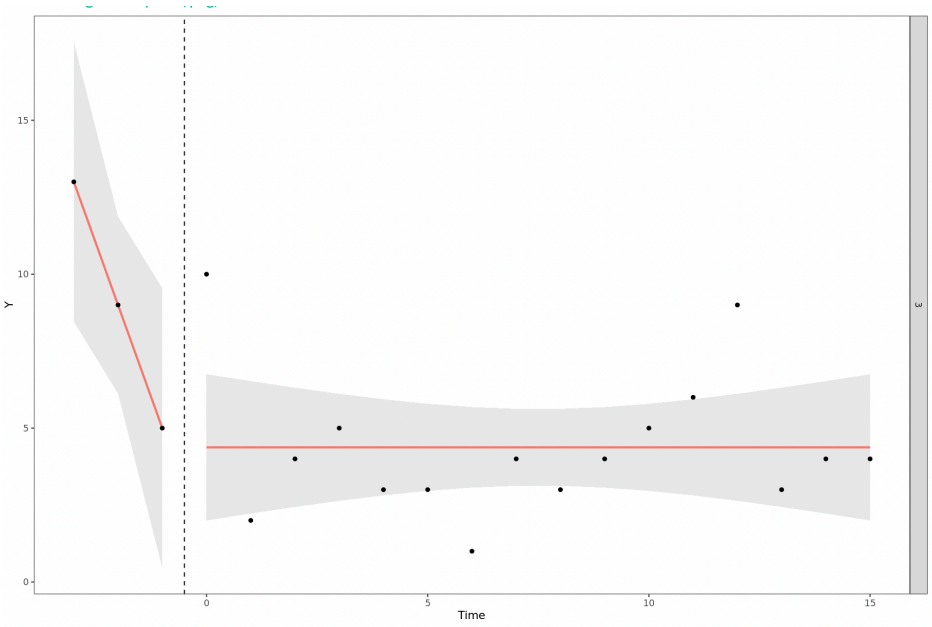


Subject	Comment
3	

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	3	(Intercept)	1	3.566	0.28	0.783
2	3	Time	-4	1.651	-2.423	0.029
3	3	Phase1	3.375	3.736	0.903	0.381
4	3	Time:Phase1	4	1.656	2.416	0.029

This patient, female, 24 years old, gets a diagnosis of *unspecified anxiety disorder* during the phase of assessment. According to Bischoff et al. (2020) 's criteria for determining GAD7 reliable change, it is possible to affirm that the patient experienced clinically significant change given that the first registered score on GAD7 during the assessment phase was 13 and the last registered score during the treatment phase was 4.

After conducting the one-level analysis it can be noticed this SCED shows a steep slope already in the baseline condition, where the expected score decreases by -4 points per time unit ($\hat{\beta}_1 = -4\,, p < 0.05$). On the contrary, the treatment phase doesn't seem to have itself an immediate impact in decreasing the symptoms which, however, remain steady (around $Y_i = 4$) on the time trend, with a slope in the treatment phase being almost equal to zero ($\hat{\beta}_3 = 4\,, p < 0.05$).

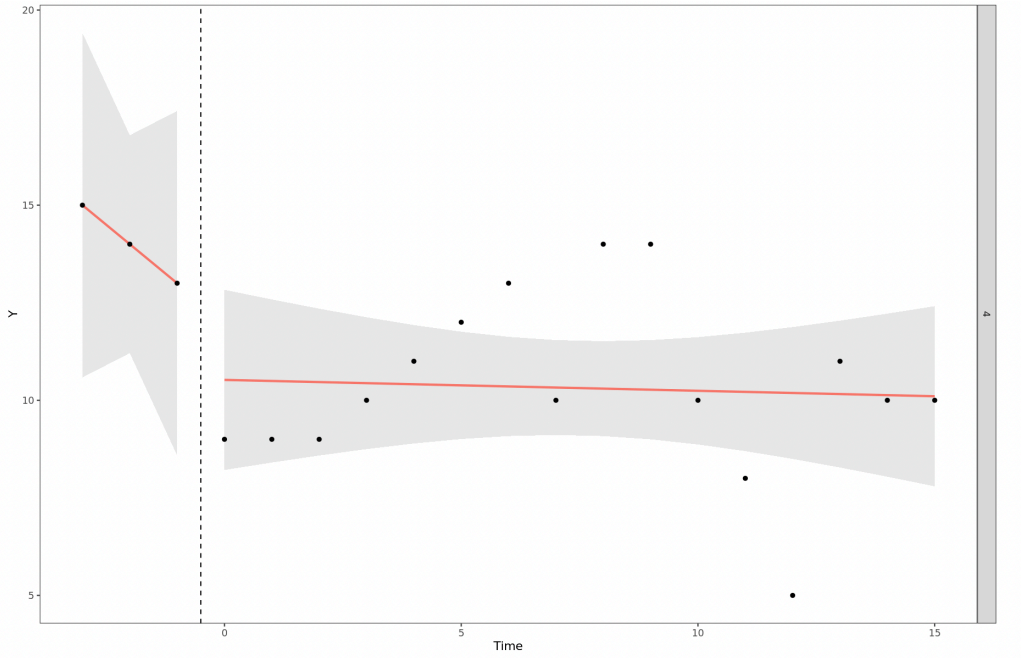


Subject	Comment
4	

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	4	(Intercept)	12	3.465	3.463	0.003
2	4	Time	-1	1.604	-0.623	0.542
3	4	Phase1	-1.478	3.63	-0.407	0.69
4	4	Time:Phase1	0.972	1.609	0.604	0.555

This patient, female, 29 years old, gets a diagnosis of a *generalized anxiety disorder* during the phase of assessment. According to Bischoff et al. (2020) 's criteria for determining GAD7 reliable change, it is <u>not</u> possible to affirm that the patient experienced clinically significant change given that the first registered score on GAD7 during the assessment phase was 15 and the last registered score during the treatment phase was 10.

Accordingly, given the high variability of the measurements, all the coefficients of the one-level analysis are not significant. Thus, it is not possible to to draw any conclusion regarding the efficacy of the treatment.

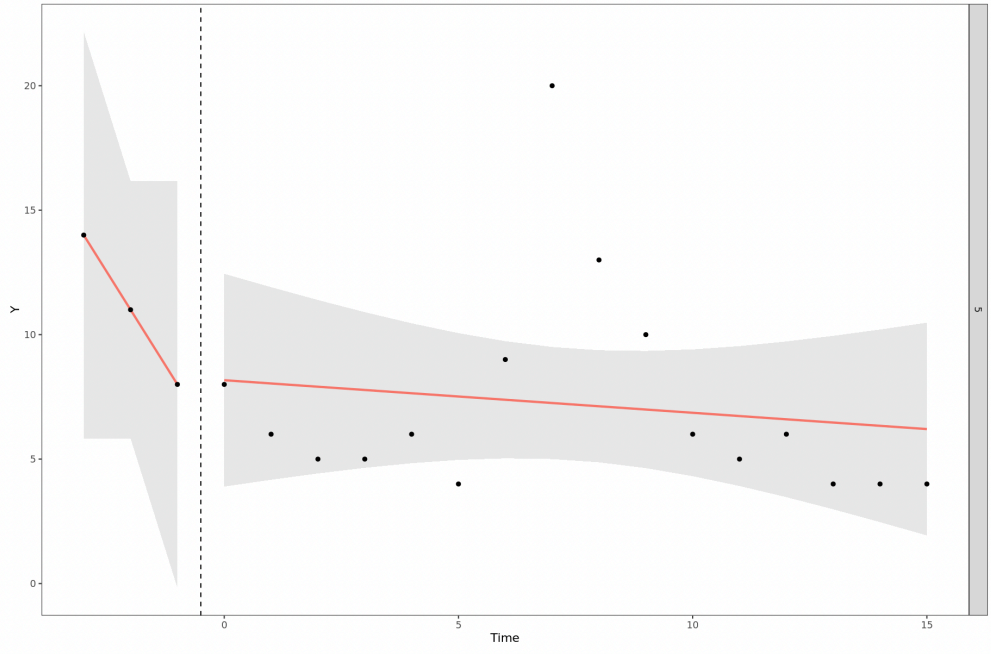


Subject	Comment
5	

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	5	(Intercept)	5	6.416	0.779	0.448
2	5	Time	-3	2.97	-1.01	0.328
3	5	Phase1	3.169	6.722	0.471	0.644
4	5	Time:Phase1	2.869	2.979	0.963	0.351

This patient, female, 53 years old, gets a diagnosis of *generalized anxiety disorder* during the phase of assessment. According to Bischoff et al. (2020) 's criteria for determining GAD7 reliable change, it is possible to affirm that the patient experienced clinically significant change given that the first registered score on GAD7 during the assessment phase was 14 and the last registered score during the treatment phase was 4.

Accordingly, given the high variability of the scores, all the coefficients of the one-level analysis are not significant. In particular, three outliers registered over the cutoff of 10 might have consistently influenced the regression analysis. Therefore, it is not possible to draw any conclusion regarding the efficacy of the treatment.

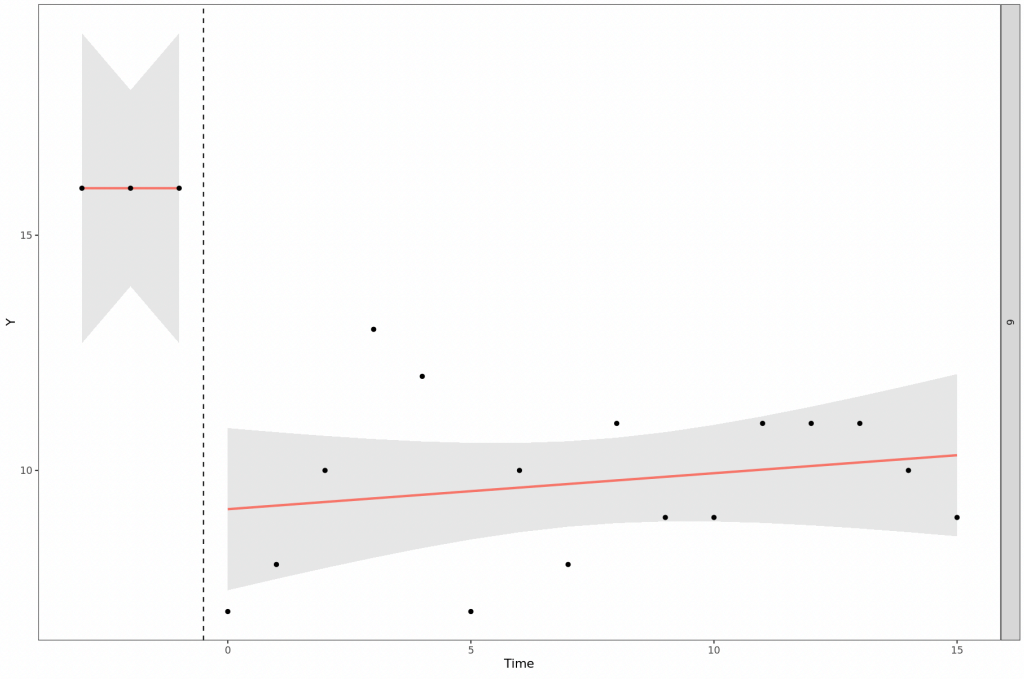


Subject	Comment
6	

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	6	(Intercept)	16	2.587	6.186	0
2	6	Time	0	1.197	0	1
3	6	Phase1	-6.824	2.71	-2.518	0.024
4	6	Time:Phase1	0.076	1.201	0.064	0.95

This patient, female, 20 years old gets a diagnosis of a *generalized anxiety disorder* during the phase of assessment together with depressive symptoms. According to Bischoff et al. (2020) 's criteria for determining GAD7 reliable change, it is possible to affirm that the patient experienced clinically significant change given that the first registered score on GAD7 during the assessment phase was 16 and the last registered score during the treatment phase was 9.

After conducting the one-level analysis it can be noticed that the estimated intercept $\beta_0 = 16$ is statistically significant. This implies that if the baseline continued, the patient would have had a GAD7 score of 16 at the start of the treatment. Interestingly, the other regression's coefficients' involving the change in time, Time ($\beta_1 = 0$) and Time:Phase1 ($\beta_3 = 0.076$) are very far from significance. It looks like time and the treatment effect on the time doesn't account for any change in the symptomatology. However, the Phase1 regressor coefficient ($\hat{\beta}_2 = -6.824$) is strongly significant meaning that, for this particular patient, the start of the intervention held a statistically significant immediate effect on her symptoms. Indeed, it appears that after the big initial improvement, the patient's anxiety level remains the same during the whole treatment phase (and the follow-up).



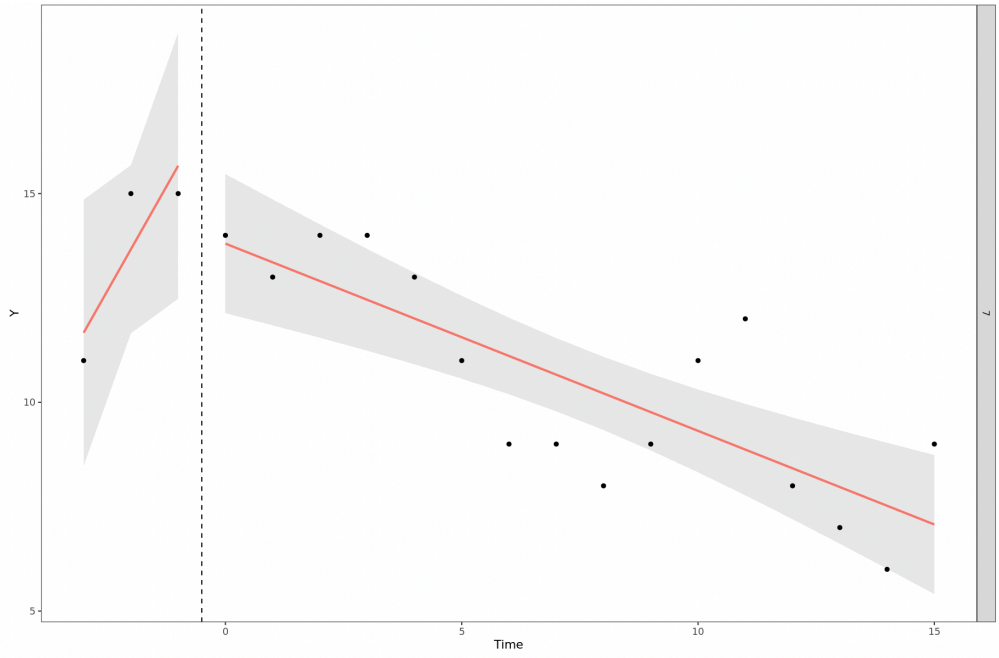
Subject	Comment

7

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	7	(Intercept)	17.667	2.501	7.064	0
2	7	Time	2	1.158	1.728	0.105
3	7	Phase1	-3.865	2.62	-1.475	0.161
4	7	Time:Phase1	-2.449	1.161	-2.109	0.052

This patient, male, 70 years old gets a diagnosis of a *generalized anxiety disorder* during the phase of assessment. According to Bischoff et al. (2020) 's criteria for determining GAD7 reliable change, it is possible to affirm that the patient experienced clinically significant change given that the last registered score on GAD7 during the assessment phase was 15 and the last registered score during the treatment phase was 9.

After conducting the one-level analysis it can be noticed that the estimated intercept β_0 = 17.667 is statistically significant. This implies that if the baseline continued, the patient would have had a GAD7 score of 17.667 at the start of the treatment. If on one hand, Time and Phase coefficients (β_1= 2 and β_2 = –3.865) are not statistically significant, on the other hand, the estimated interaction of Time and Phase1 (β_3= –2.449) is statistically significant (p-value= 0.052) meaning that the treatment effect on the time trend is statistically significantly different from zero.

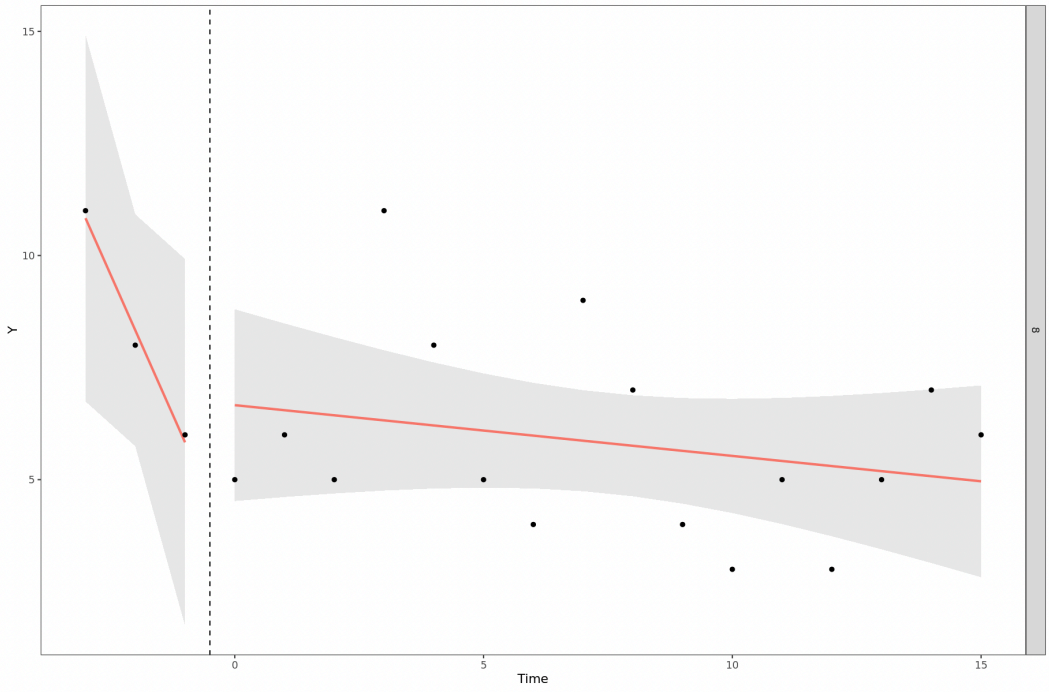


Subject	Comment
8	

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	8	(Intercept)	3.333	3.21	1.038	0.316
2	8	Time	-2.5	1.486	-1.682	0.113
3	8	Phase1	3.328	3.363	0.99	0.338
4	8	Time:Phase1	2.387	1.49	1.601	0.13

This patient, female, 21 years old, gets a diagnosis of a *generalized anxiety disorder* during the phase of assessment. According to Bischoff et al. (2020) 's criteria for determining GAD7 reliable change, it is not possible to affirm that the patient experienced clinically significant change given that the first registered score on GAD7 during the assessment phase was 11 and the last registered score during the treatment phase was 6.

Accordingly, given the high variability of the scores, all the coefficients of the one-level analysis are not significant. Therefore, it is not possible to draw any conclusion regarding the efficacy of the treatment.

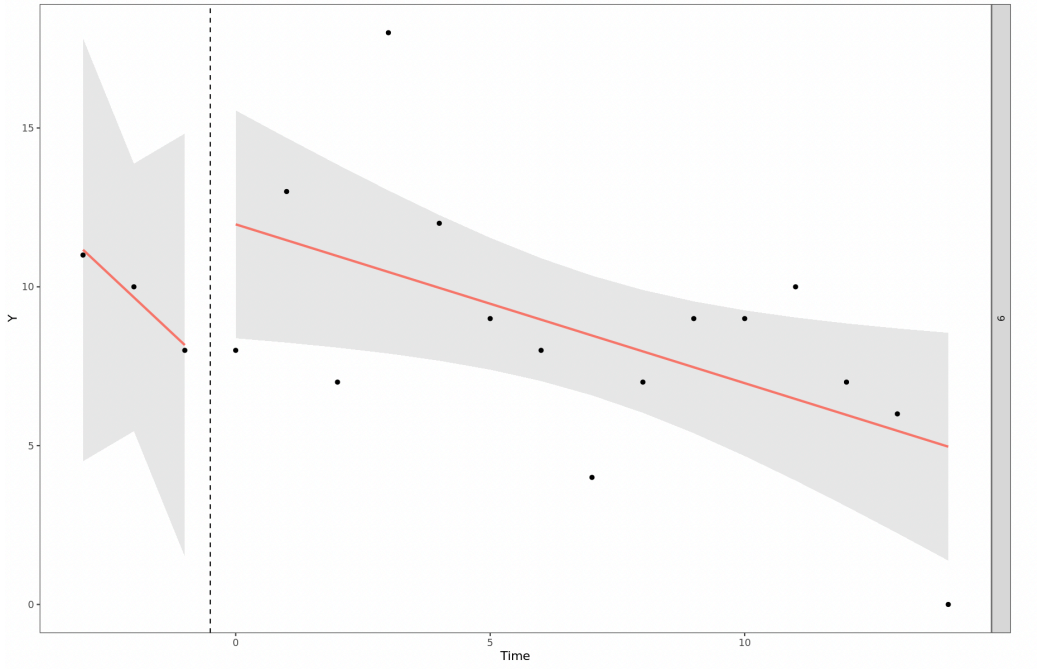


Subject	Comment
9	

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	9	(Intercept)	6.667	5.195	1.283	0.22
2	9	Time	-1.5	2.405	-0.624	0.543
3	9	Phase1	5.3	5.457	0.971	0.348
4	9	Time:Phase1	1	2.413	0.414	0.685

This patient, female, 20 years old, gets a diagnosis of a *generalized anxiety disorder* during the phase of assessment. According to Bischoff et al. (2020)'s criteria for determining GAD7 reliable change, it is possible to affirm that the patient experienced clinically significant change given that the first registered score on GAD7 during the assessment phase was 11 and the last registered score during the treatment phase was 0.

Unfortunately, given the high variability of the scores, all the coefficients of the one-level analysis are not significant. Therefore, it is not possible to draw any conclusion regarding the efficacy of the treatment.



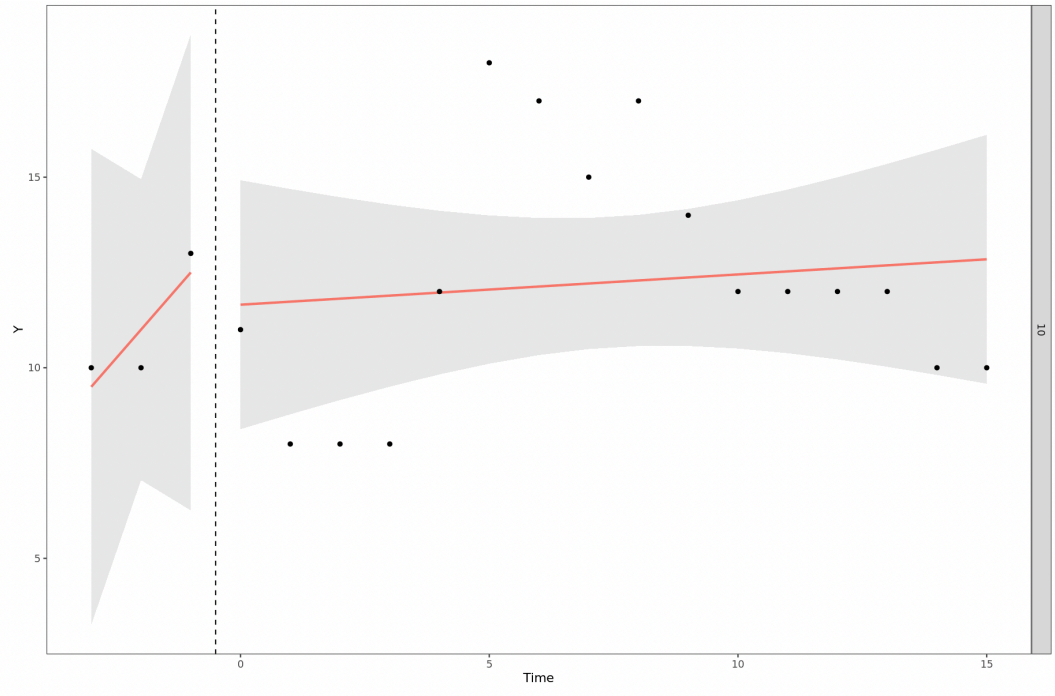
Subject	Comment

10

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	10	(Intercept)	14	4.9	2.857	0.012
2	10	Time	1.5	2.268	0.661	0.518
3	10	Phase1	-2.346	5.134	-0.457	0.654
4	10	Time:Phase1	-1.421	2.275	-0.624	0.542

This patient, female, 52 years old, gets a diagnosis of *unspecified anxiety disorder* during the phase of assessment. According to Bischoff et al. (2020) 's criteria for determining GAD7 reliable change, it is not possible to affirm that the patient experienced clinically significant change given that the first registered score on GAD7 during the assessment phase was 10 and the last registered score during the treatment phase was 10.

After conducting the one-level analysis it can be noticed that the estimated intercept $\beta_0 = 14$ is statistically significant. This implies that if the baseline continued, the patient would have had a GAD7 score of 14 on the symptomatology at the start of the treatment.
Unfortunately, all the other coefficients are not significant, hence it is not possible to draw any conclusion regarding the efficacy of the treatment. Besides, from the visual inspection, it can be observed that the slope in the intervention phase is increasing, indicating that the treatment wasn't having an effect on symptoms reduction.



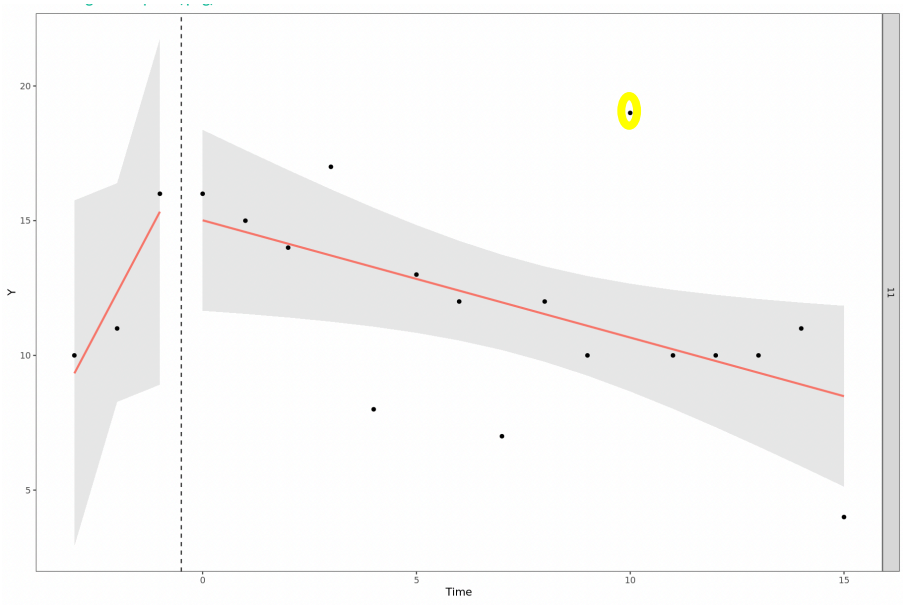
Subject	Comment
11	

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	11	(Intercept)	18.333	5.039	3.638	0.002
2	11	Time	3	2.333	1.286	0.218
3	11	Phase1	-3.319	5.28	-0.629	0.539
4	11	Time:Phase1	-3.435	2.34	-1.468	0.163

This patient, male, 57 years old, gets a diagnosis of *unspecified anxiety disorder* during the phase of assessment. According to Bischoff et al. (2020)'s criteria for determining GAD7 reliable change, it is possible to affirm that the patient experienced clinically significant change given that the first registered score on GAD7 during the assessment phase was 10 and the last registered score during the treatment phase was 4.

Unfortunately, given the high variability of the registered scores, all the coefficients of the one-level analysis are not significant. Therefore, it is not possible to draw any conclusion regarding the efficacy of the treatment.

However, the visual inspection shows a clear symptom decrease in the intervention phase. For this reason, we tried to delete the outlier point highlighted in yellow. After reconducting the one-level analysis, the coefficients Phase1 and Time:Phase were significant ($\beta_2 = 10{,}6$ $\beta_3 = -3{,}5$, $p < 0.05$). This shows that, even if the treatment didn't play a role in the symptom's reduction at the beginning of the intervention, it was responsible for the change on the time trend.



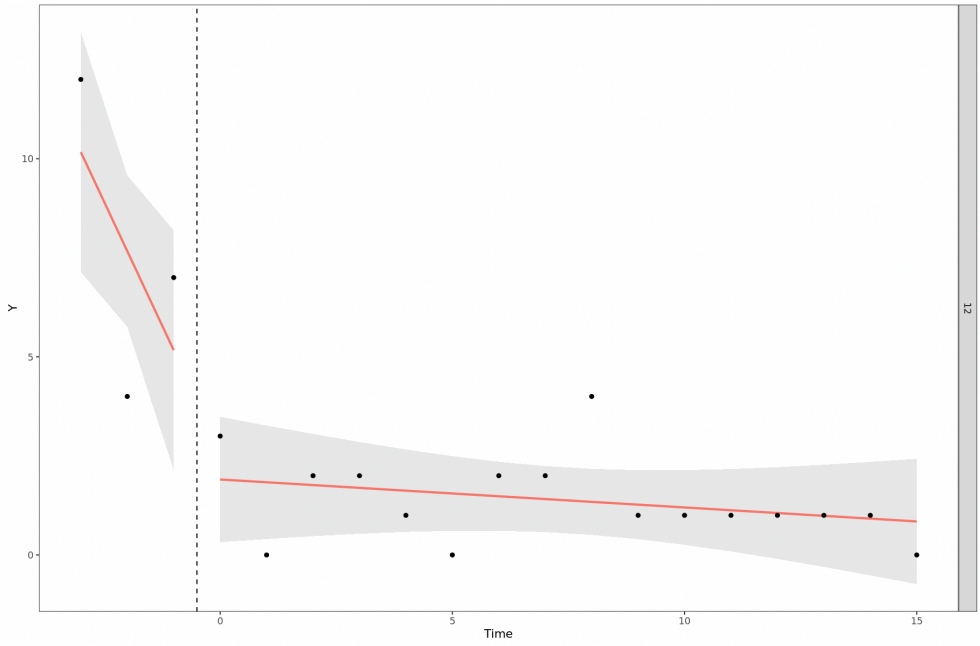
Subject	Comment
12	

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	12	(Intercept)	2.667	2.374	1.123	0.279
2	12	Time	-2.5	1.099	-2.275	0.038
3	12	Phase1	-0.762	2.487	-0.306	0.763
4	12	Time:Phase1	2.429	1.102	2.204	0.044

This patient, female, 28 years old, gets a diagnosis of *unspecified anxiety disorder* during the phase of assessment. According to Bischoff et al. (2020)'s criteria for determining GAD7 reliable change, it is possible to affirm that the patient experienced a significant change given that the first registered score on GAD7 during the assessment phase was 12 and the last registered score during the treatment phase was 0.

After conducting the one-level analysis it is worth noticing that the estimated intercept $\beta_0 = 2.667$ is not statistically significant whereas the Time coefficient is significant ($\beta_1 = -2.5$, $p < 0.05$). This shows that the score of symptoms, without the start of the treatment, would potentially decrease by 2.5 with each additional time unit.
Interestingly, Phase 1 ($\beta_2 = -0.762$) is not significant, meaning that the intervention doesn't have a statistically significant immediate effect. On the contrary, the estimated interaction of Time and Phase1 is statistically significant ($\beta_3 = 2.429$, $p < 0.05$) meaning that the treatment effect on the time trend is statistically different from zero.

From the visual inspection, it can be observed that, differently from most of the other cases, this specific SCED show less outlier points; this could support the hypothesis that outlier measurements play an important role in the quality of the results.

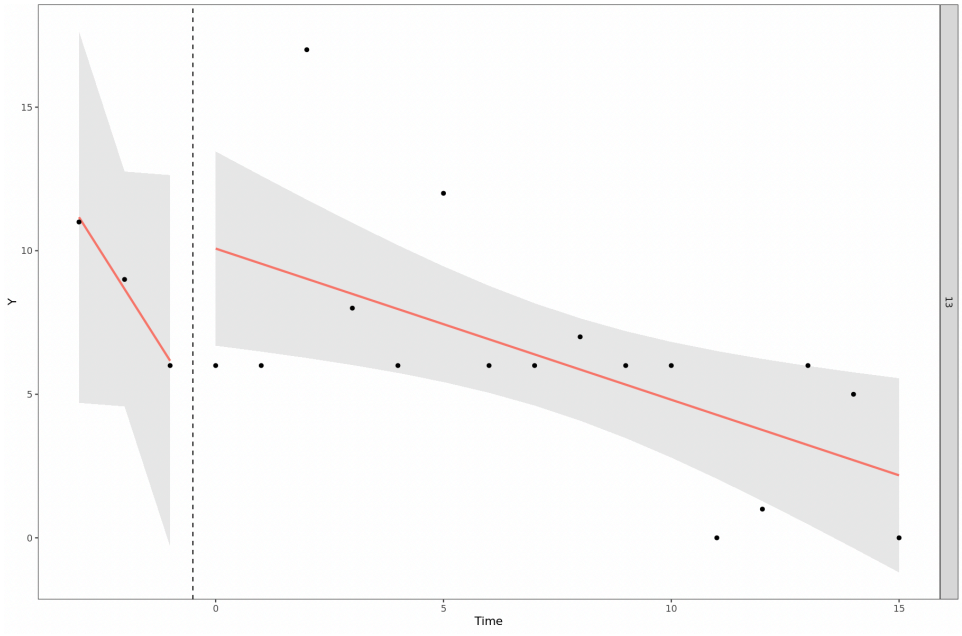


Subject	Comment
13	

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	13	(Intercept)	3.667	5.077	0.722	0.481
2	13	Time	-2.5	2.35	-1.064	0.304
3	13	Phase1	6.407	5.319	1.205	0.247
4	13	Time:Phase1	1.974	2.357	0.837	0.416

This patient, male, 40 years old, gets a diagnosis of *unspecified anxiety disorder* during the phase of assessment. According to Bischoff et al. (2020) 's criteria for determining GAD7 reliable change, it is possible to affirm that the patient experienced clinically significant change given that the first registered score on GAD7 during the assessment phase was 11 and the last registered score during the treatment phase was 0.

Unfortunately, given the high variability of the scores, all the coefficients of the one-level analysis are not significant. Therefore, it is not possible to draw any conclusion regarding the efficacy of the treatment. Two outliers registered over the cutoff of 10 might have consistently influenced the regression analysis.



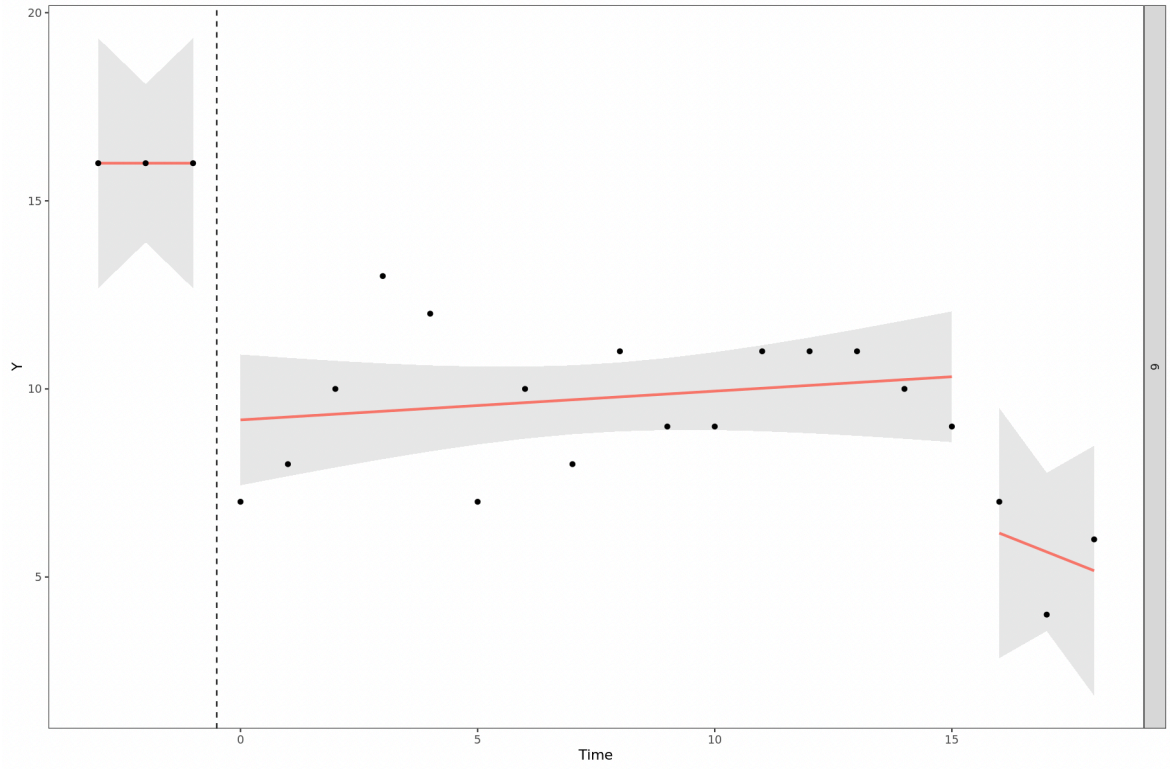
D. Effect sizes predictions

Name	Time after treatment	Effect size	Lower limit	Upper limit
1	15	-5.12	-6.01	-4.24
2	15	-5.60	-6.37	-4.84
3	15	-4.62	-5.41	-3.84
4	15	-3.69	-4.34	-3.03
5	15	-3.81	-5.06	-2.56
6	15	-6.25	-6.74	-5.76
7	15	-3.23	-4.02	-2.43
8	15	-2.52	-3.19	-1.85
9	15	-1.20	-2.41	0.01
10	15	1.25	0.32	2.18
11	15	-0.58	-1.74	0.57
12	15	-6.29	-6.81	-5.77
13	15	-2.54	-3.76	-1.33

E. Follow-up

Output of the one-level analysis computed through the MultiSCED (Declercq et al., 2020) and considering the follow-up scores.

	Name	Regressor	Coefficient	Standard Error	t-value	p-value
1	6	(Intercept)	16	2.623	6.1	0
2	6	Time	0	1.214	0	1
3	6	Phase1	-6.824	2.748	-2.483	0.024
4	6	Phase2	-1.833	20.831	-0.088	0.931
5	6	Time:Phase1	0.076	1.218	0.063	0.951
6	6	Time:Phase2	-0.5	1.717	-0.291	0.775



Plot of the two-level analysis outcome displayed by the MultiSCED (Declercq et al., 2020), considering the follow-up scores.