Challenges to Scaling a Youth Violence Reduction Intervention: Evidence from a Field Experiment and Qualitative Data*

Sarah Cattan Pulkit Bajpai Julian Edbrooke-Childs Emily Stapley

Monica Costa-Dias Jenna Jacob Emily Goodacre Angelika Labno

Vaishnavi Dhas Emily Orchard Erin Mckeaveney Grizelda Khaling

Anoushka Kapoor Ashley New Imran Rasul

September 2026

*We gratefully acknowledge financial support from the ESRC (ES/H021221/1, ES/Y002563/1) and Impact Acceleration Account, the Stone Centre at UCL, the Nuffield Foundation (SFS /FR-000024397) and the Youth Endowment Fund (YEF). We thank all the individuals and organizations that have been part of the study, including the practitioners, Single Points of Contacts, clinical leads, trainers, and service managers in Local Authorities, and all the young people who participated. We are indebted to Karla Goodman, Isabelle Gregory and Ben Byrne and other project team members at LIIA and the VRU for their collaboration and to Florence Kroll for her support for evaluation and chairing of the programme steering group. We thank the team at the Youth Endowment Fund, in particular Daniel Acquah, Sarah Fullick and Rain Sherlock. We are grateful to Navya Malik, Emily McDougal, Innamana Pettyll, and young people on the Young People Advisory Group. We thank Tasnim Nodee, Harriet Olorenshaw and Yann Calvo-Lopez for excellent research support during the pilot trial and whose efforts contributed to the project transitioning to the scale-up phase. We received ethics approval from UCL (5114/014, 5115/013). The pilot and efficacy trials are registered with a pre-analysis plan (AEARCTR0009611, ISRCTN69049265). This publication is based on research originally funded and commissioned by the YEF. We acknowledge the YEF's support and involvement in the design and delivery of the original study. The analysis and views presented here are those of the authors. All errors remain our own. Cattan: IFS [sarah_c@ifs.org.uk]; Bajpai: IFS [pulkit.bajpai@ifs.org.uk]; Edbrooke-Childs: AFC [Julian.childs@annafreud.org]; Stapley: AFC [Emily.Stapley@annafreud.org]; Costa-Dias: Bristol and IFS [monica_d@ifs.org.uk]; Jacob: AFC [jenna.jacob@annafreud.org]; Goodacre: AFC [ejg69@cam.ac.uk]; Labno: AFC [angelika.labno@annafreud.org]; Dhas: IFS [vaishnavi.dhas@ifs.org.uk]; Orchard: AFC [emily.orchard@annafreud.org]; Mckeaveney: AFC [mckeaveneyerin@gmail.com]; Khaling: AFC [yanakhaling@gmail.com]; Kapoor: AFC [anoushka.kappor@annafreud.org]; New: AFC [ashley.new.21@alumni.ucl.ac.uk]; Rasul: UCL and IFS [i.rasul@ucl.ac.uk].

Abstract

Concerns around the exposure of young people to violence lie at the top of the UK policy agenda. We evaluate a flagship violence-reduction intervention in London, known as *Your Choice*, that targets adolescents at risk of exposure to violent crime. The intervention trains youth practitioners in CBT, who deliver these techniques to adolescents over 12-18 weeks. Delivery is embedded within pre-existing services that London Local Authorities are statutorily required to provide, so accounting for constraints to scale-up from the outset. Many challenges to retaining voltage at-scale relate to system-wide factors that are hard to measure quantitatively. We therefore combine qualitative and quantitative evidence to understand intervention impacts moving from an internal pilot to a full scale efficacy trial. While the pilot showed the promise of *Your Choice*, the intervention faced challenges to scale. We document four reasons for this: (i) initial pilot participants differ from those recruited at scale-up; (ii) uncertainty over future funding and staff turnover; (iii) binding workload capacity constraints for practitioners; (iv) a rise in per participant intervention costs. We derive implications for future evaluations of similar interventions in public services that operate near full capacity, to help build the evidence base on how to overcome these fundamental sources of voltage drops. *JEL: I20, J13.*

1 Introduction

Adolescence is a period of life when individuals are at high risk of being exposed to violence. Whether in the form of being a perpetrator, a victim or witness to violence, such exposure can worsen outcomes related to mental health, education and employment [Hjalmarsson 2008, Aizer and Doyle 2015, Owens 2017, Eren and Mocan 2021].[1]

Concerns around exposure of young people to violence lie at the top of the policy agenda in the UK, our setting. While official statistics document a steady decline in overall youth crime, violent crime among youth is increasing. Panel A in Figure 1 shows trends in youth crime from official statistics: violent crime against the person increased by 13% between 2014 and 2024, while over the same period many other crime types, such as theft and burglary, declined [YJB 2024]. Panel B shows that by 2024 there were 12,000 proven violent offences committed by young persons, more than three times the incidence of the next most prominent crime type, constituting a third of all offenses committed by young people. Many of these are serious crimes, with the majority of youth with custodial sentences having committed a violent crime. These trends are reflected in rising annual social costs of youth crime: by 2018/9 these reached £1.3 billion in England and Wales, an increase of more than 50% over four years [Irwin-Rogers *et al.* 2020].

Against this backdrop, the London Mayoral Office established the Violence Reduction Unit (VRU) in 2019 to develop early intervention strategies to reduce youth exposure to violence. Following an acute year for youth violence in London, with 30 teenage homicides in 2021, the VRU launched a new initiative to tackle the issue. Our study is a close collaboration with the VRU and other London-wide stakeholders to evaluate this flagship intervention. The program, known as *Your Choice* (YC), is backed by a £10mn budget and three-year time horizon, signalling a big push top-down commitment for a step-change in service provision to vulnerable youth.

The intervention targets adolescents aged 11-17 assessed to be at risk of exposure to violent crime and motivated by evidence that young people involved in violence could benefit from therapeutic support [Gaffney *et al.* 2021]. The intervention trains professional staff employed by Local Authorities (LAs) – youth practitioners, who work with vulnerable adolescents – in CBT-techniques. Over 12-18 weeks, eligible young are intended to people meet their YC-trained practitioner in three weekly sessions of 45-60 minutes each. As with many CBT-based interventions for young people [Lipsey *et al.* 2007, Beck 2011], YC aims to help them develop an understanding of the thinking processes and choices that precede violent behavior, and provide mechanisms to help them slow down decision-making and consider situations and the consequences of their actions [Ross and Nisbett 2011, Heller *et al.* 2017].[2]

[1] Advances in brain imagery suggest adolescence is a period of significant reorganization of the brain, with the control centre (the prefrontal cortex) being the last part to develop. Neurological studies have documented that risk-related behavior is correlated with prefrontal activity [Leijenhorst *et al.* 2010, Schonberg *et al.* 2012], helping to understand spikes in risk taking behavior during adolescence [Steinberg 2008].

[2] A key concept of CBT is that an individual's thoughts, feelings, and behaviors are interlinked. *Your Choice*

The delivery and evaluation of the intervention are both firmly embedded within pre-existing services that LAs are statutorily required to provide to vulnerable young people. By constraining the evaluation to take place within the real-world practicalities of service delivery, our study anticipates challenges for scale-up. By collecting information on sources of potential voltage drops at both the internal pilot and efficacy trial stages of the evaluation, our study is designed to maximize learning about the *effectiveness* of the intervention at-scale, rather than focusing on its *efficacy* under ideal conditions, results not generalizable at-scale. As such, our study falls into the realm of 'practical programs' [Lipsey 1999], taking to heart option-C thinking, namely generating evidence at pilot and scale-up stages to inform an understanding of the key constraints interventions face at-scale [List 2024].[3]

Youth services dealing with at-risk adolescents are organized into teams of practitioners, with each practitioner belonging to a single team. LAs identified teams that could potentially be trained in YC. Within each LA, teams were either randomized into treatment, namely all practitioners in the team were trained in YC, or left as controls. Treated teams delivered the intervention to all eligible adolescents assigned to them. Following pre-existing practices, young people would be referred to the practitioner team with the most capacity at the time of their assessment. This generates randomness in the assignment of adolescents to treated and control teams, enabling us to estimate treatment effects on adolescent outcomes from them being assigned to a YC-trained team relative to a non-trained team.

Our evaluations at pilot and scale-up combine quantitative data – from administrative records, baseline and endline surveys to adolescents and practitioners – and qualitative data collected from young people, practitioners and key individuals involved in implementation in each LA. The qualitative data plays a prominent role in our analysis because many common causes of voltage drops relate to system-wide factors – the emergence of complexities and uncertainties at-scale, the gradual loss of buy-in etc. Many of these are hard to measure quantitatively, yet remain critical for understanding challenges to scale.

The pilot took place from June 2022 to March 2023. The aim was to train practitioners in delivering YC, to inform the feasibility of delivering the intervention, the fidelity of delivery within

combines elements of traditional CBT approaches, where thinking and behavior are key targets for change, with a more recent and broader approach to change the relationship an individual has with their inner world. *Your Choice* also emphasizes the importance of the therapeutic relationship. Hence the focus on intense contact between young people and their practitioners to build trusting relationships, increasingly considered a pre-requisite for effective behavioral change [Beck 2011, Wenzel 2021, Dumornay *et al.* 2022].

[3] Lipsey [1999] represents an early emphasis on this distinction in the area of juvenile crime. He describes efficacy as the demonstration that a program can produce effects under ideal circumstances, while effectiveness refers to evidence that a program produces effects under conditions of routine implementation. He highlights typical characteristics of programs that are implemented for research purposes, focus solely on efficacy and differ from routine practices: special training for service providers, specialized selection criteria for service recipients, detailed treatment protocols, monitoring and supervision of treatment by researchers, and lower attrition prior to completion of service. As challenges at-scale are not well anticipated, pilots can then exhibit more promise than is achieved at scale-up.

the experimental design, and the feasibility of constructing a London-wide data infrastructure to evaluate the intervention. We then had a period of taking learnings from the pilot before moving to scale from December 2023 to June 2025. Our evaluation corresponds to a multi-site study, covering 30 out of 32 LAs in London across pilot and scale-up stages.

Our primary quantitative outcome relates to conduct problems as measured by the *Strengths and Difficulties Questionnaire* (SDQ) [Goodman *et al.* 1997]. Our pre-analysis plan pre-specified our primary outcome to be an indicator for whether the young person scored in the high/very high range of conduct problems. This was our primary outcome because the association between conduct problems and disruptive behavior including violent and non-violent delinquency is well-established [Spaan *et al.* 2023]. Our pre-analysis plan also specified our secondary outcomes. Each links to the theory of change – either because they reveal mechanisms through which CBT-tools delivered by practitioners might reduce adolescents' exposure to violence including emotional regulation, or they capture intended outcomes of the intervention related to prosocial behavior and wellbeing.

The pilot is based on 193 adolescents across 43 teams in 25 LAs. The evidence from the pilot showed promise of *Your Choice*. While no single piece of evidence is compelling given the small-sample, the combination of quantitative and qualitative evidence suggested: (i) the incidence of conduct problems being in the high/very high range was reduced by 11pp, corresponding to a 33% fall from the baseline mean in controls; (ii) a high fidelity of implementation with the research design; (iii) the intervention operating through the intended theory of change, with trusting relationships between adolescents and practitioners being built and the delivery of CBT-techniques in treated sessions.

The scale-up is based on 1465 adolescents across 122 teams in 30 LAs. Despite being powered to detect the same impacts as in the pilot, YC failed to meet its initial promise when implemented at-scale. On the primary outcome, the likelihood of high/very high conduct problems actually *rose* slightly by 2.8pp. While this is not statistically significant ($p = .328$), the 95% confidence interval rules out any fall larger than 2.9pp corresponding to 11% of the baseline mean in controls, and an impact that is significantly different from the point estimate reduction observed in the pilot ($p = .000$).

Many interventions that show promise in small-scale pilots fail to deliver results at-scale [List 2022, 2024] including in the area of juvenile crime [Lipsey 1999, Stevenson 2023]. The remainder of our analysis combines the quantitative and qualitative evidence to drill down to understand challenges to scaling. We follow the typology described in List [2026] to examine four challenges: (i) sample unrepresentativeness, where participants at-scale differ in characteristics or motivation from those involved in the pilot; (ii) situational unrepresentativeness, where conditions related to the intervention differ at-scale from those in the pilot; (iii) supply-side constraints, where organizational uncertainties, complexities and spillovers only emerge at-scale; (iv) increasing marginal costs of delivery.

On sample unrepresentativeness, we find the baseline pilot sample to be more vulnerable in terms of conduct problems and self-assessed risks than those recruited at scale-up. To assess whether this compositional change explains the voltage drop, we reweight the scale-up sample to match baseline means among the pilot sample using the entropy balancing approach of Hainmueller [2012]. Estimating treatment effects on our primary outcome in this reweighted sample, we find that for those treated with YC, the likelihood of being in the high/very high range of conduct problems falls by 11pp relative to controls – almost identical to the impact achieved in the pilot. The loss in voltage can thus nearly all be quantitatively explained by the change in composition of young people recruited at scale-up. An implication is that the intervention impacts are then likely to be heterogenous, and larger for more vulnerable youth. We confirm this to be so in the scale-up sample. For example, for our primary outcome, the intervention reduces the likelihood of being in the high/very high range of conduct problems significantly more for those that were in this range to begin with at baseline: the differential impact across groups is 13.6pp ($p = .059$). Overall, the evidence points to the intervention having greater impact on the primary and some secondary outcomes for those more vulnerable to begin with, a group that were recruited into the pilot to a great extent than those recruited at scale-up.

This change can occur between phases because the latter phase captures the regular flow of young people assessed to be at medium/high risk of exposure to violence, while the pilot sample better reflects vulnerable youth assigned to local authorities and services ready and willing to engage with the initial pilot implementation of YC.

A second class of challenges arise due to situational unrepresentativeness, where conditions related to the intervention differ at-scale from those in the more controlled environment of a pilot. A central design feature of YC is that it is adaptable to the needs of specific youth that practitioners are working with. While these features continued to be seen as a key to its success at-scale in the qualitative data, our evidence reveals two costs of such adaptability. The first is that tailoring delivery led to a shortfall in the number of sessions that young people received relative to the intended intensity of treatment – although this shortfall occurred over both study phases and therefore is not key to understanding the voltage drop at-scale. Second, adaptable delivery is a double-edged sword in that there is no single formula or recipe to be passed on to ensure successful delivery to future adolescents. This is especially problematic for an intervention operating within real world constraints of adolescent services that rely on the adoption of a train-the-trainer model to retain voltage, where clinical leads train the next wave of practitioners in YC.

A third class of challenges to scaling are supply-side constraints, where organizational uncertainties and complexities only emerge at-scale. Such challenges relate to system-wide factors, and so are hard to measure quantitatively. Our qualitative data sheds light on some key supply-side constraints. The first is bottom-up buy-in. The qualitative data suggests the intervention gained practitioner buy-in because it was perceived to add to their toolkit, providing opportunities for their professional development, and bring structure to their work and the young people they

worked with. The second issue highlighted by the qualitative data relates to a key uncertainty only realized at-scale: staff turnover. Staff turnover means that young people's involvement in YC could end or be disrupted if their practitioner left, or services could lose entire practitioner teams. Turnover of clinical leads in LAs led to particularly acute strains on voltage because of the consequent reductions in clinical supervision, and a reduced ability to train incoming practitioners. A third complexity that emerged at-scale relates to workloads: practitioners felt that if the intensive work around YC was to continue, they would need to have their workloads reduced. Operating at scale increased the demands on capacity as practitioners were required to recruit and deliver the intervention to more young people than in the pilot. Practitioners consistently reported the challenge of delivering the intense level of contact for a young person receiving YC while maintaining their usual case load.

A fourth challenge relates to rising costs at-scale. Cost data collected from a few local authorities tentatively suggests the emergence of diseconomies of scale are a potential final cause of voltage drops: the average cost per participant almost doubled from around £9000 at the pilot phase to over £17,000 at scale-up. The reasons relate to the challenges to scaling related to situational unrepresentativeness and supply-side constraints described earlier, specifically: (i) higher costs related to the clinical supervision of practitioners – that partly reflects staff turnover among senior leads in LAs preventing gains from specialization to be fully realized; (ii) costs incurred as local authorities provided resources to enable young people to continue with elements of YC to sustain its impact, despite the lack of future funding guarantees.

The implementation science literature has focused on fidelity issues as being the key challenge to sustaining voltage [Kilbourne *et al.* 2007, Weiss *et al.* 2014, Gottfredson *et al.* 2015, Al-Ubaydli *et al.* 2019], and specifically among interventions to reduce juvenile recidivism, meta-analysis suggests the majority of voltage losses are due to administrative fidelity [Lipsey 1999]. Throughout we do not find evidence that a lack of fidelity with the research design is the primary cause for voltage drops: the number of participants recruited into scale-up was in line with requirements suggested by our post-pilot power calculations, attrition rates fell moving to scale-up given the lessons learned from the pilot and mitigation strategies put into place, samples remained mostly well balanced, and the research design was well adhered to at-scale. This finding in itself is no small achievement: this is the first time these LAs had been involved in any large-scale evaluation, while operating at full capacity, and with well-established norms on how to deliver services to vulnerable young people. At a concluding stage of analysis, we draw together our evidence on the challenges *Your Choice* faced to scale, deriving implications for option-C testing in similar contexts, namely, to go beyond A/B testing and build evaluation designs that shed quantitative light on what might ameliorate sources of voltage drops that we identify in these kinds of contexts.

Our work contributes to an established literature in psychology evaluating CBT-based interventions, albeit in pilot samples, using non-experimental methods, and reaching adolescents either via schools, families or correction facilities [Lipsey *et al.* 2007, Wilson *et al.* 2015], and a nascent

literature in economics experimentally examining the impacts of such interventions at-scale.

Meta-analyses in criminal psychology have highlighted the promise of CBT-based approaches targeting vulnerable populations [Del Vecchio and O'Leary 2004, Saini 2009], but with variation in intervention effectiveness depending on the underlying risks faced by targeted groups or the intensity of delivery [Lipsey 1999, Landenberger and Lipsey 2005, Lipsey *et al.* 2007, Koehler *et al.* 2013]. Closer to the population targeted by YC, the evidence suggests CBT-based interventions targeting young people with conduct disorders can lead to large improvements in externalizing behaviors such as conduct problems [Durlak *et al.* 2001, Riise *et al.* 2021, Wichmann *et al.* 2023, de Haan *et al.* 2024]. In the economics literature, a recent wave of studies have evaluated CBT-based interventions at-scale, many taking place in Chicago [Heller *et al.* 2017, Bhatt *et al.* 2024, Abdul-Razzak *et al.* 2025], and documented interventions targeting very high risk males can lead to long-term reductions in involvement in violent crime.[4]

None of this earlier work studies the issue from the perspective of the science of scaling, that lies at the heart of our analysis, and enables us to dig deeper into *why* interventions to reduce adolescents' exposure to violence may not scale [Lipsey 1999, Stevenson 2023]. The existing evidence around CBT-interventions already hints at features relevant for effectiveness at-scale: impacts are significantly larger in pilot programs than in interventions delivered within the context of existing practice, in high-fidelity studies, when delivered to high-risk and older participants, and when more intense treatment is delivered [Bennett and Gibbons 2000, Landenberger and Lipsey 2005, Lipsey *et al.* 2007, Koehler *et al.* 2013]. There is also evidence that the *Becoming a Man Program* in Chicago worked well initially but failed to scale because of the kinds of double-edged sword considerations that arise from adaptable intervention designs as was the case for *Your Choice* [Heller *et al.* 2017, Bhatt *et al.* 2021].[5]

Our evaluation is a collaboration with stakeholders responsible for a flagship London-wide intervention at-scale, with practicalities of service delivery informing the design of the pilot and scale-up.

[4]Heller *et al.* [2017] evaluate a school-based intervention in high-crime neighborhoods of Chicago: *Becoming a Man*. They report results from three at-scale RCTs ($n > 2,000$). The intervention involved CBT-sessions, delivered to groups of 15 students weekly by non-specialists. They document a 50% fall in arrests for violent crime and sustained gains in high school graduation. Bhatt *et al.* [2024] evaluate *READI* in Chicago, an intervention targeting individuals at very high risk of being involved in gun crime. It delivered a CBT-informed group intervention together with an 18-month subsidy to prevent illegal work. Findings suggest large improvements in gun violence but lower impacts on other violent crimes. Abdul-Razzak *et al.* [2025] evaluate another intervention in Chicago, *Choose to Change*, that engages with hard-to-reach young people through mentors ($n > 2,000$). Eligibles were aged 13-18 and were actively affiliated with gangs or at high risk of gang engagement, on probation, previously found guilty of weapons offenses, disruptive in school, or victims of or witnesses to traumatic violence. As with YC, the intervention involved building trusting relationships with mentors. CBT was delivered weekly in groups by specialist therapists over 12-16 weeks. The evaluation provided some evidence of reduced arrests for violent offences, but with no evidence that arrests for other offences were impacted. Outside the US, Blattman *et al.* [2017, 2023] document how the CBT-therapy targeted to groups of former male combatants in Liberia, reduced engagement in violence, with such effects being more likely to persist when therapy was combined with cash grants.

[5]Bhatt *et al.* [2021] suggest reasons for this related to the lack of skilled counselors recruited at scale-up, and the inherent flexibility of the intervention meaning it could be tailored to specific needs, but as a result, there was no unique recipe for how the intervention worked. We discuss both issues later in our context.

This representativeness of the situation is key to scaling [Levitt and List 2007, Al-Ubaydli *et al.* 2021], and reflects a well recognized feature in implementation science, that complex infrastructures are required to support programs [Braithwaite *et al.* 2018, Fagan *et al.* 2019]. Comparing the evidence across our study phases provides a nuanced picture of multiple challenges to scaling. We advocate for more combined use of quantitative and qualitative data collection exercises, given that many challenges relate to system-wide factors that are hard to measure quantitatively and sometimes unrelated to the behavior of participants or delivery agents. Such mixed methods approaches provide one avenue through which the science of scaling can help improve understanding of how to ameliorate voltage drops in the context of public services operating with little slack.

Section 2 describes our study context, the *Your Choice* intervention, our research design and data sources. Sections 3 and 4 present findings from the pilot and scale-up. Section 5 delves into understanding the challenges to scaling the intervention. Section 6 discusses implications for future A/B/C testing designs for similar interventions. Appendix A presents additional results, Appendix B describes additional qualitative evidence and Appendix C shows survey modules.

2 Context, Intervention and Data

2.1 Context

London is made up of 32 Local Authorities (LAs). Each is obliged to deliver statutory services to vulnerable youth at risk of exposure to violence. These services vary across LAs in terminology and exact practices, but core services supporting young people can be categorized into *Children Looked After*, *Youth Justice*, and *Adolescent Safeguarding*. Services are delivered by professional staff employed by LAs – youth practitioners. Practitioners are organized into teams within each service, with each practitioner belonging to a single team and there can be more than one team per youth service (depending on the size of the LA). Young people can be referred to LAs via schools, the police, contact with other LA services or self-referral. Young people are then assessed by a multi-agency child exploitation panel, and identified to be either of low, medium or high risk of harm, that includes committing or being a victim of violence. They are assigned to the practitioner team with the most capacity at the time of their risk assessment. Young people then work with the same practitioner within a team.

An important feature of the context – common to many public services in large cities targeting vulnerable populations – is that youth services operate at full capacity, with a relatively fixed workforce and little slack. This will be relevant when assessing challenges to scaling.

2.2 Your Choice

The *Your Choice* (YC) intervention is a London-wide partnership initiative of the London Innovation and Improvement Alliance (LIIA) and London's Violence Reduction Unit (VRU). We collaborated with key stakeholders and administrators from the London Mayoral office responsible for the intervention at-scale, including LIIA and the VRU. Both supported the design, implementation and evaluation of YC. To support the effectiveness of the intervention, a Young People's Action Group (YPAG) was also established to co-produce materials for the evaluation, and provide feedback to the evaluation team to ensure: (i) young peoples' perspective remained central to the study; (ii) the intervention design and materials were appropriate for the hard-to-reach and vulnerable youth targeted, most of whom are from minority backgrounds.[6]

Intervention Design YC targets those aged 11-17 and evaluated to be at medium/high risk of committing violent acts or be vulnerable to being a victim of violence outside of their families. Clinical practitioners within LIIA, with a background of working in youth justice, designed the intervention. YC trains youth practitioners to use CBT-techniques to support the young people they work with. The intervention was designed to be delivered over 12 weeks, with adolescents taking part in three weekly sessions of 45-60 minutes each with YC-trained youth practitioners.

YC has three core elements. First, intense delivery of sessions was intended to build trusting adolescent-practitioner relationships. To further this aim, the intervention was flexibly delivered in accessible settings such as youth centres, cafes and gyms. Practitioners were also encouraged to tailor delivery to the individual needs of young people. For example, tools could be adapted to suit needs including speech, language and communication, autism spectrum disorder, and learning disabilities. The YC intervention is built on the hypothesis that these practitioners are better at establishing trusting relationships with them than clinical psychologists.[7]

Second, sessions were informed by the use of CBT tools by practitioners. As with many CBT-based interventions, YC emphasizes the need for helping emotional and cognitive regulation through developing emotional awareness and regulation, social and problem solving skills. This reflects the needs of the target cohort, who, by virtue of their developmental stage and exposure to trauma and adversity, are often more likely to respond from emotional or physiological states

[6] The VRU was set up by the London Mayoral Office in 2019. It partners with communities, Local Authorities, the NHS, police, schools and charities to understand the causes of violence and prevent it using early intervention approaches. The London Innovation and Improvement Alliance is the Association of London Directors of Children's Services sector-led improvement partnership. LIIA serves to share best practices to find London-wide solutions that positively impact children's services and the lives of young people. The YPAG comprised nine members (aged 16-24) from diverse backgrounds and lived experience of mental health care, social care, or youth justice. The group convened four times to help develop questionnaires, discuss recruitment strategies and ensure materials accounted for the racial and ethnic diversity of the targeted cohort.

[7] The planned intensity of YC matched those of other CBT-interventions [Riise *et al.* 2021] and flexibility in the location of sessions helped address a concern raised with interventions delivered in schools, that the most at-risk youth are typically disengaged from such institutions [Abdul-Razzak *et al.* 2025].

before engaging with cognitive processes. Helping young people with this process is thus key to preventing violent behavior, that is often a result of hypersensitive fight or flight responses, automatic distorted thinking and impulsive decision making [Ross and Nisbett 2011, Heller *et al.* 2017]. YC aims to: (i) support the young person develop an understanding of the emotional, psychosomatic and thinking processes and responses that precede such behaviors; (ii) equip them with coping strategies through emotional regulation and understanding cognitive processes; (iii) empower them to disrupt unwanted patterns of behavior that might impact their future goals.

Third, there is clinical supervision of practitioners. Motivated by evidence suggesting that repeated exposure to vicarious trauma can inhibit practitioners' ability to provide high-quality care [Quitangon 2019], such supervisions provide a space for practitioners to talk about cases with a colleague with psychological expertise, helping them manage the emotional load of cases.

The control business-as-usual condition for young persons varies across LAs. While all those aged 11-17 evaluated to be at medium/high risk of committing acts of violence are assigned a youth practitioner, their engagement with practitioners differs in each of the core ingredients of YC: (i) the control condition involves less structured and intense support than YC; (ii) it is not geared towards using CBT tools; (iii) it does not necessarily include clinical supervision to practitioners – limiting scope for practitioners to learn from each other, and to deal with the pressures of hearing traumatic life stories from young people.[8]

Theory of Change The theory of change underlying YC is that it aims to: (i) reduce emotional and behavioral difficulties for targeted youth; (ii) improve their attitudes and activities around education, employment and training; (iii) reduce their exposure to violence and improve their well-being. The theory proposes that the intervention achieves this through the following mechanisms: (i) increases the engagement of adolescents with practitioners and other support networks; (ii) increases young persons self-understanding and provides them self-management strategies; (iii) allows young persons to progress towards their goals and improve their prosocial identity.[9]

These aims and mechanisms underlie our data collection. Our primary outcome relates to conduct problems measured at baseline and endline by the *Strengths and Difficulties Questionnaire*

[8] In the UK, the National Institute for Health and Clinical Excellence recommends CBT for the management of mental health difficulties, including with younger people dealing with psychological and behavioral issues. However, given a lack of service availability, complicated referral processes, poor information, cultural barriers and stigma, it is well recognized that young people with the highest risk factors are least likely to access such support in a clinical setting [Crenna-Jennings and Hutchinson 2020].

[9] These aims and mechanisms are common to other CBT-informed interventions targeting violent behavior, as well as a wider literature documenting that non-cognitive skills, including self-control and forward looking behavior, predict long run criminal activity [Heckman *et al.* 2006]. For instance, *Becoming and Man* combined role playing with activities to promote reflection and introspection, identify negative emotions and thoughts, work through alternative perspectives, and improve social behavior and attitudes towards education and training [Heller *et al.* 2017]. Goal setting is also routinely used to promote engagement by giving participants a sense of purpose, and to structure skill building and behavioral change [Lipsey *et al.* 2007, Heller *et al.* 2017, Abdul-Razzah *et al.* 2025]. The triggering of prosocial behaviors among young offenders has been documented to reduce antisocial behaviors in the context of restorative justice interventions [Shem-Tov *et al.* 2024, Adukia *et al.* 2025].

(SDQ) [Goodman *et al.* 1997]. Table C1 details the questions used to construct a 0-10 score for the conduct problems subscale. Conduct problems are directly tied to the theory of change in relation to cognitive restructuring and impulse control [Ross and Nisbett 2011, Heller *et al.* 2017]. Our pre-analysis plan also pre-specified our secondary outcomes. Each links to the theory of change – either because they reveal mechanisms through which CBT-tools delivered by practitioners might reduce adolescents' exposure to violence including emotional regulation, or they capture intended outcomes of the intervention related to prosocial behavior and wellbeing.

Intervention Delivery In each LA, a lead in children's service provision was designated as a Single Point of Contact (SPoC) for YC. They were responsible for ensuring the intervention was resourced and evaluated in line with the research design. The SPoC identified two senior colleagues in their LA to deliver YC: (i) a Clinical Lead: a clinician with expertise in delivering CBT, responsible for the quality of the intervention's clinical provision; (ii) a YC trainer to teach practitioners CBT-techniques. Practitioners received five days of training in CBT theory and skills via a train-the-trainer model and monthly clinical supervision with the Clinical Lead. The intervention is thus designed to be delivered entirely within pre-existing structures and constraints of statutory adolescent services. By leveraging the existing workforce in LAs, the intervention faces less risk of voltage drops as a result of supply-side issues such as later having to engage with the relevant administrative staff in each LA and gain their buy-in.[10]

2.3 Research Design

LAs were asked to identify teams that could potentially be trained in YC. Randomization was performed at the team level within each LA, by the evaluation team. Teams were either randomized into treatment, namely all practitioners in the team were trained in YC, or left as controls. Treated teams were tasked to deliver the intervention to all eligible adolescents assigned to them. Control teams continued to deliver business-as-usual support to adolescents.

Teams randomized into control were told they would be later offered to be trained in YC. Hence, an assumption underpinning our design is that practitioners in control teams do not change behavior towards adolescents assigned to them in anticipation of later being trained in YC.

Once adolescents were assessed to be of medium/high risk of exposure to violence and thus eligible for YC, following pre-existing practices, they would be referred to the practitioner team with the most capacity at the time of their assessment. This mechanism aimed to generate randomness in the assignment of adolescents across treated and control teams. Randomization of individual adolescents was deemed infeasible given the way youth services normally operate, risking interfering with the delivery of statutory services. A key aspect of LAs objection to individual

[10] Further information and tools about delivering the intervention was made available on the YC page of the LIIA website (https://www.liia.london/your-choice). Further details of both phases are in Cattan *et al.* [2023, 2025].

randomization was they could not guarantee the availability of allocated spaces in any given team given: (i) limited capacity of teams; (ii) capacity fluctuates over time.[11]

The assignment mechanism of adolescents to teams followed the same rules during the pilot and scale-up, for treatment and controls.

Two points are key in relation to the design validity.

First, the design leverages the usual assignment process of adolescents to teams, based on workload considerations in a system working at capacity. This guarantees as good-as-random assignment, whereby the assignment of adolescents to teams is independent of their characteristics and potential gains from treatment. Second, the recruitment of adolescents is required to be independent of their treatment assignment. In the pilot and scale-up, adolescents were recruited to participate after they were assigned to teams. Any systematic differences in recruitment across groups could therefore invalidate randomization. To minimize the risk of selective recruitment, LAs were instructed not to disclose to the young person which type of support they would receive until they had consented and completed the baseline questionnaire. Naturally, practitioners were not blind to the allocation since they knew whether they had been trained in YC.[12]

Violations of either the capacity-based assignment rule, or of the balanced recruitment condition, amount to non-compliance and can lead to selective enrolment.

2.4 Data Sources

Our evaluation combines information from five sources: (i) administrative records providing background information on adolescents; (ii) practitioner completed session forms, recording details of each session with a young person (in treated and control groups); (iii) baseline and endline surveys to adolescents, where endline took place around 20 weeks after recruitment and hence after the completion of YC-sessions; (iv) baseline and endline surveys to practitioners; (v) qualitative data collected from young people, practitioners and key individuals implementing YC in each LA. The data collection tools followed nearly exactly the same design at pilot and scale-up stages.[13]

[11] An exception to this assignment rule is applied to adolescents returning to LA services after a brief interruption. Returning adolescents would be assigned to the same team they previously worked with. As their assignment to teams precedes YC and the treatment assignment of teams is randomized, characteristics of returning adolescents should not correlate to whether they are assigned to treated or control team.

[12] After their assignment to a practitioner, the practitioner informed the adolescent about the opportunity to participate in the study. They showed them the recruitment video and information sheet, and discussed its content alongside a privacy notice. For those aged 11-15, the practitioner also provided their parent(s)/guardian(s) with the parents' information sheet and had similar discussions with them. Their parent/guardian was also required to sign the parental consent form. Adolescents received a £10 (£25) voucher upon completion of their baseline (endline) questionnaire.

[13] Baseline and endline surveys of young people were completed during an adolescent-practitioner session. Surveys were built with a function to have questions read out loud by a young person's voice. In rare cases (such an adolescent with acute special educational needs), the practitioner may have helped the young person complete the survey.

2.4.1 Quantitative Data

Primary Outcome Motivated by the theory of change, our primary outcome relates to conduct problems measured by the *Strengths and Difficulties Questionnaire* (SDQ) [Goodman *et al.* 1997]. Conduct problems are directly tied to the main aim of the intervention, as stated in the theory of change, in relation to cognitive restructuring and impulse control. The association between the SDQ conduct problems subscale and disruptive behavior is well-established [Spaan *et al.* 2023]. Relative to actual measures of criminal/violent activity, the SDQ conduct problem subscale has the advantage that it is readily available during our study period, while engagement in violent crime might take longer to occur.[14]

While the SDQ subscales can be used as continuous scores, it is also possible to categorize scores into ranges. SDQ score cut points are chosen such that 80% of children in a population-based UK survey score in the average range, 10% score in the borderline range, and 10% score in the extreme high/very high range. Our pre-analysis plan specified our primary outcome to be an indicator for whether the young person scored in the high/very high range of conduct problems.

Secondary Outcomes We also pre-specified our secondary outcomes. Each links to the theory of change – either because they capture intended outcomes of the intervention related to prosocial behavior and wellbeing, or they reveal mechanisms through which CBT-tools delivered by practitioners might reduce adolescents' exposure to violence including emotional regulation.

Our first set of secondary outcomes are measured from additional subscales from the SDQ at baseline and endline: (i) hyperactivity; (ii) emotional difficulties; (iii) peer difficulties; (iv) prosocial behavior. Table C1 details the questions used to construct a 0-10 score for each subscale. We again classify whether individuals score in the high/very high range for each dimension of conduct relative to the population of young people. The Appendix discusses validation of the data quality around the SDQ-derived outcomes.

We also pre-specified another four secondary outcomes based on established scales related to self-regulation, social connectedness, prosociality and wellbeing. These outcomes all relate to the theory of change. The scales used are: (i) the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire – Adolescent Short Form; (ii) the Social Connectedness Scale – Revised: this measures the young person's resilience through strengthened sources of social support at home, school and elsewhere; (iii) the Prosocial Identity Scale; (iv) the Short Warwick-Edinburgh Mental Well-being Scale [Ng Fat *et al.* 2017]. Scales (i) and (iv) are only measured at endline in the pilot and scale-up; Scale (iii) is only measured in the scale-up. These scales have been validated on

[14] A majority of at-scale randomized studies on CBT interventions use criminal activity as the primary outcome because in most cases, these interventions are targeted at offenders [Heller *et al.* 2017, Bhatt *et al.* 2024, Adbdul-Razzah *et al.* 2025]. Our pre-analysis plan did specify a verifiable measure of offending behavior as a secondary outcome. This requires us to the link our data to administrative records from the Police National Computer. Given the time required for such linkages to be permitted, we leave the study of this outcome to future work.

diverse populations.

Our final outcome relates to self- and practitioner-reported perceptions of young person's safety – a holistic measure of exposure to violence. These are measured through a new instrument developed by the project and evaluation teams, that we refer to as the *Checkpoint* scale [Goodman *et al.* 2022]. Table C2 shows the questions used in the *Checkpoint* scale – these include subjective measures of the propensity to commit violent acts or vulnerability to being a victim of violence. The Appendix discusses the context-specific motivation for developing this scale, and given its importance as a holistic measure of exposure to violence, also discusses its validation.

2.4.2 Qualitative Data

We complement the quantitative analysis with qualitative data assembled from interviews with: (i) adolescents: during the pilot we interviewed those assigned to treated and control groups, while in the scale-up interviews focused on treated adolescents; (ii) practitioners delivering YC; (iii) YC trainers; (iv) other senior leads in LAs responsible for the delivery of adolescent services. Staff in all LAs were contacted and invited to express an interest to be interviewed. We selected staff to ensure SPoCs, clinical leads, and senior managers were recruited. These senior staff were recruited from 10 LAs, and practitioners from 16 LAs. Young people could consent to interview when consenting for the overall study, and we invited young people to provide coverage of a range of LAs. We also monitored the demographic characteristics and purposively sampled groups that were underrepresented in initial interviews.[15]

We draw on validated methods to establish the trustworthiness of the data [Yardley 2000] and categorize transcript extracts [Ritchie and Spencer 1994, Bruan and Clarke 2006]. Drawing on the Framework Approach [Ritchie and Spencer 1994], we organized the content of transcripts into pre-defined categories, such as items related to the theory of change. To do so, we conducted reflexive thematic analysis, identifying codes emerging from participants' responses through an iterative process [Bruan and Clarke 2006]. The Appendix further details data collection protocols, including the handling of safeguarding concerns arising during interviews.

Combining qualitative and quantitative data sources provides complementary perspectives on the fidelity of the evaluation and whether and how this changed from the pilot to scale-up. More-

[15] Interviews were conducted by phone or video call. They were audio recorded and transcribed verbatim. To preserve confidentiality, we have not included identifiers for interview quotes. We worked with practitioners and the YPAG to adjust how we were communicating with young people with special educational needs when inviting them for interview, and how we conducted interviews. Participation in the qualitative interviews was voluntary: young people completed a consent form with their practitioner and staff completed a consent form online or verbally. Consent was verbally confirmed at start of the interview. At the start of their interview and throughout, participants were reminded that they could take a break or stop at any point and to only talk about what they felt comfortable with. Two researchers conducted each interview and both participated in a debrief with another senior team member after the interview to discuss safeguarding concerns, reflect on the interview, and discuss any impact on the researcher. Interviews on average lasted for 28 minutes (young people), 48 minutes (practitioners), and 43 minutes (other leads from LAs). After the interview, young people received a £10 voucher.

over, such mixed methods approaches can be critical for understanding challenges to scaling because some potential causes of voltage drops relate to system-wide factors – the emergence of complexities and uncertainties at-scale, the gradual loss of buy-in etc. and many of these are hard to measure quantitatively. Our qualitative data sheds light on potential sources of voltage drops relating to: (i) the nature of the intervention, captured by perceptions of treatment gains from young people and their practitioners; (ii) the implementation process at-scale – including top-down perceptions from YC trainers and other leads in LAs, and bottom up perceptions from practitioners; (iii) the emergence of uncertainties/complexities at scale, and its perceived sustainability and spillovers within a system already operating at full capacity.

2.5 Timeline

The study timeline comprises an internal pilot and full scale efficacy trial. Figure 2 shows the participant flow and how this corresponds to the timing of data collection. This is identical at pilot and scale-up stages. The consistency in these design features allows us to treat the pilot as internal, so maximizing learnings from the pilot to the scale-up. The pilot had two phases. In December 2021, five LAs each put forward two potential teams to be trained in YC, one of which was randomly selected by the evaluation team to actually be trained. Training was extended to additional teams during the Spring of 2022. In the majority of cases, LAs again put forward two teams and the selection of treated and control teams was randomized within LA. Baseline data was collected from June 2022 to December 2022. Endline data was collected from December 2022 to March 2023. All the subsequent analysis of the pilot focuses on those LAs that put forward at least two teams that could then be randomly assigned to treatment or control. Other LAs still started to implement YC but they non-randomly selected a team for YC during the pilot – those data play no role in the analysis.

We had a period of taking learnings from the pilot to modify data collection protocols before moving to scale-up. At the scale-up stage, LAs nominated two or more untrained teams for randomization. The evaluation team conducted the randomization. Training was rolled out to randomly selected teams between June 2023 and April 2024. Teams that participated during the pilot and randomized into YC continued to deliver the intervention. Adolescents were recruited from August 2023 to December 2024, with endline data collected from December 2023 to June 2025. Qualitative data collection took place from November 2023 to March 2025. To meet participant numbers, in January 2025 it was decided to extend the recruitment period until the end of February 2025 for a select number of LAs.

3 Pilot

The pilot analysis is based on 193 adolescents assigned to a team randomly selected to either deliver YC or held as controls. The sample covers 43 teams across 25 LAs. 21 teams were assigned to treatment and 114 (59%) of adolescents were assigned to treated teams. Panel A of Table A1 shows the LA services that teams in the pilot are from: the majority of recruited youth are assigned to teams in *Children Looked After* services.

3.1 Balance

We first present descriptive evidence on our sample. This shows a target population grappling with emotional and behavioral difficulties and that is highly exposed to violence.

Table 1 shows baseline characteristics of adolescents related to their: (i) demographics (Panel A); (ii) primary needs (Panel B); (iii) activities (Panel C). We report the mean in treatment and control, their difference, and the p-value on a t-test of equality of means. Panel A shows most youth are male (70%), there is an even split of younger and older adolescents (within the eligible age range of 11-17) and most are from minority backgrounds: among controls, 25% are White, 27% are Black, 14% are Asian and 11% are mixed race. Individuals have high rates of special educational needs (23%), disability (44%) and of being in social care or looked after by their LA (18%). Panel B shows the most common primary need assessed relates to family dysfunction (20% in controls), followed by socially unacceptable behaviors (14%). Panel C shows the activities of adolescents. As expected, the majority report being in school/college (60% in controls), with similar shares of controls being in work and training (7%). For 13 out of 18 characteristics we cannot reject equality between treatment and controls at conventional levels of significance.[16]

Table 2 shows balance on the SDQ subscales. Among controls, 37% of youth have a score indicating conduct problems in the high/very high range. 37% score in the high/very high range for hyperactivity, 29% do so for emotional problems, 38% for peer problems, and 21% score in the low/very low range for prosocial behaviors. The ranges on the SDQ measures are calibrated so that 10% of the youth population are expected to lie at these extremes: our sample is thus between two and four times as likely to be in the extreme range of problems for each. The final row reports balance on another secondary outcome at baseline: the social connectedness scale. Table A2 shows balance on each subcomponent of this scale, that relates to networks of support.

[16] Among all 11-17 year olds in London in 2024/5, 12% are classified as receiving special educational needs support, and 5.6% receive intensive support, often for a disability. Far less than 1% of children under the age of 18 are in the care of their LA. We describe the qualitative evidence below, but highlight how the quantitative data match descriptions from practitioners of the young people they work with. The qualitative data highlighted there was: (i) disengagement of adolescents from the 'system', referring to education, employment and training; (ii) a mistrust of professionals; (iii) feelings of criminal victimization and trauma; (iv) complex needs or conflict in the home. This reinforces the notion that trusting relationships with practitioners need to build to drive behavioral change [Abdul-Razzak *et al.* 2025].

The samples are balanced for each SDQ-based outcome, both on indicators of whether problems lie in extreme ranges or the continuous score, and for the social connectedness scale.

Table 3 focuses on our final secondary outcome: assessed risks of engaging in risky behaviors, as developed in the *Checkpoint* scale. Panel A shows adolescent's self-reports of being likely/very likely to be exposed to various risks in the next month. Even over a short horizon, the average young person in our sample faces two (out of five) risks, and treatment and controls are balanced on their risk exposure ($p = .576$). Around 10% of adolescents view it as likely/very likely they will cause damage in a public place, get into trouble with the police, physically hurt someone, or that someone else will involve them in such behaviors. 20% report that they are likely/very likely to witness a violent crime, and 6% report being likely/very to be a victim of violent crime. Panel B uses data from the *Checkpoint* scale administered in the baseline survey to practitioners. Over the same time horizon, practitioners perceive adolescents to be exposed to three out of the five risks – so a higher degree of exposure than adolescents themselves report. This aggregate risk assessment is balanced ($p = .161$), and the sample is balanced for five out the six of the perceived risks.

The results from Tables 1 to 3 suggest that at an aggregate level, randomization protocols appear to have been followed, consistent with the assumptions underlying the research design of: (i) a capacity-based assignment rule; (ii) balanced recruitment into the trial.

3.2 Attrition

Of the 193 adolescents in the pilot, 96 are tracked to endline – an attrition rate of 50%. Table A3 examines correlates of attrition: (i) unconditionally, attrition is not statistically different between treatment and controls (Column 1); (ii) this remains the case conditional on adolescent characteristics (Column 2); (iii) there is evidence of heterogeneous attrition on observables (Column 3): the joint F-test of interactions of characteristics with treatment is significant ($F = 24.7$). This is driven by higher rates of attrition from the treated group for white, black and mixed race/ethnicity adolescents, and those assessed to be at lower risk on the *Checkpoint* scale by practitioners. We detail below the robustness of the pilot findings to accounting for selective attrition, and the mitigation strategies put into place after the pilot that successfully ameliorated the extent and selective nature of attrition at scale-up.

3.3 Implementation

To document the extent to which YC was delivered by practitioners with the intended intensity, we examine evidence from session forms from adolescent-practitioner meetings. A caveat is that in the pilot there were challenges in collecting this data: we received 778 sessions data, 638 (140) from treated (control) practitioners, with 65% (35%) of treated (control) practitioners submitting at least one session form. Taking this data as purely exploratory, it suggests: (i) far more sessions took

place under YC than for controls; (ii) for 65% of controls there is no report of any session, while this falls to 37% for treated adolescents; (iii) the average number of reported sessions (conditional on at least one session) is 8.9 (5) for treated (controls) – 27% of adolescents in YC received 10 or more sessions (including zeroes), far lower than the intended 36 sessions over 12 weeks; (iv) among treated adolescents, when sessions took place there is strong evidence of the planned curricula of YC being delivered: 94% of reported sessions involved some content related to the YC curriculum, 55% involved the use of CBT-tools, 31% focused on relationship building, and 20% involved working towards goals.

3.4 Results

To estimate treatment effects of being assigned to a YC team relative to a control team, we use the following specification for young person i in local authority l:

$$y_{il} = \alpha + \delta y_{0il} + \beta T_{il} + u_{il}, \tag{1}$$

where y_{il} is the outcome of interest, y_{0il} is the outcome at baseline (where available), T_{il} is a dummy equal to one if adolescent i is assigned to a practitioner team trained in *Your Choice* and zero otherwise, and u_{il} is an error term. As treatment assignment does not necessarily imply that young person i receives YC-informed sessions, we interpret the β estimates as ITT parameters. Following Chaisemartin and Ramirez-Cuellar [2024] we report standard errors allowing for clustering at the local authority level. We report conventional p-values on the null that $\beta = 0$, and where appropriate, we also report p-values that adjust for multiple hypothesis testing using the stepdown method [Romano and Wolf 2005, Romano *et al.* 2008, Romano and Wolf 2016].

Table 4 shows the results for adolescents tracked to endline. Columns 1 to 5 show primary and secondary outcomes for the SDQ-based subscales, where we focus on whether the adolescent lies in the extreme range of each. At the foot of each column we show the mean outcome among controls at baseline for those tracked to endline.

On the primary outcome, Column 1 shows the likelihood of being in the high/very high range for conduct problems falls by 11pp (from a baseline mean of 33% in controls), corresponding to a 33% fall from the baseline mean in controls. Across secondary outcomes in Columns 2 to 5, the likelihood of being in the high/very high range of problems related to hyperactivity and peers also both fall by 8pp and 4pp respectively (corresponding to 20% and 12% falls relative to the baseline mean in controls). In contrast, the likelihood of emotional problems being in the high/very high range is 5pp higher in the treated group at endline, while the impacts on prosocial behaviors are close to zero.[17]

[17] Table A4 shows impacts on the continuous measures on each SDQ measure: the signs of the treatment effects follow those for the indicators of being in the extreme range. On the prosocial behaviors scale, treated adolescents have a marginally statistically significant .656 higher score than controls ($p = .093$), an 8% increase over the

The remaining Columns in Table 4 show treatment effects on other secondary outcomes. As with some dimensions of SDQ-measures, most point estimates move in the expected direction given the theory of change: mechanisms such as social connectedness, emotional self-regulation, and well-being all nudge higher among treated adolescents relative to controls, and self-assessed risks from the *Checkpoint* scale nudge down. At the same time, the practitioner-assessed risk *Checkpoint* scale nudges upwards.

This pattern of results is robust to: (i) controlling for characteristics that are imbalanced at baseline (Panel A, Table A5); (ii) additionally controlling for LA fixed effects, the randomization strata (Panel B, Table A5); (iii) clustering standard errors at the team level (Table A6); (iv) using bootstrapped standard errors that are clustered at the team level (Table A7). We note that our main approach yields more conservative standard errors than clustering at the team level.

Table A8 uses IPWs to adjust for selective attrition to address the concern that attriters might gain the least from YC. The result generally show treatment effects for the SDQ indicators to be larger in absolute value. For example, the likelihood of being in the high/very high range for conduct problems falls by 16pp, corresponding to a 30% fall from the baseline mean in controls, and there is a statistically significant increase in the index of mental well-being ($p = .082$).

Overall, given the small sample and high attrition, none of the estimates on primary and secondary outcomes in Table 4 are significant. While we do not expect such precision from the small-scale pilot, the direction and size of point estimates hint at the YC intervention holding some promise of positively impacting young people at risk of violence, and doing so through the intended mechanisms and on the intended outcomes as in the proposed theory of change. We now examine whether the qualitative evidence from the pilot points in the same direction.

3.5 Qualitative Evidence

To underpin the notion of promise, we consider the qualitative evidence from the pilot. This sheds light on whether the views of young people and practitioners align with YC improving outcomes of young people relative to controls, in a way consistent with the theory of change. This evidence offers insights into how the intervention was implemented, adherence to the research design, and the engagement of stakeholders. This analysis is based on 20 interviews: young people assigned to the YC treatment ($n = 5$) or control ($n = 5$), practitioners delivering YC ($n = 5$), and other LA staff involved in implementation ($n = 5$). LA staff also completed an anonymous implementation survey on the process by which eligible young people were assigned to teams in their LA ($n = 22$).

baseline mean in controls. However, we mostly find the magnitudes of impact on the continuous scores correspond to smaller percentages of the baseline mean in controls than for the high/very high range indicators. For example, the conduct problem continuous score falls by 7% of the baseline mean while the high/very high range indicator falls by 33%. This hints that YC might be more effective among those with the most severe conduct problems to begin with. We return to this issue when considering changes in the characteristics of participants at-scale as a challenge to scaling.

3.5.1 Adolescents Assigned to Treatment

Young person's descriptions of how YC was delivered aligned with its design, with sessions being held once/twice a week, lasting 30 to 60 minutes each, and taking place in a variety of accessible locations. As suggested by the session forms data, despite the far lower than intended intensity of treatment over the 12 weeks from recruitment, treated practitioners and youth both reported building trusting relationships, the utilization of CBT-based tools in sessions, and shifts in outcomes related to conduct problems and other changes in line with the theory of change. In Appendix B we summarize some of the recurring themes from interviews: (i) that YC helped build trusting adolescent-practitioner relationships; (ii) session content reflected the use of CBT-techniques; (iii) the theory of change was supported.

3.5.2 Adolescents Assigned to Control

We contrast testimonies to those from young people assigned to control teams. As expected, the qualitative data highlighted variation in support received by youth in controls across LAs. Interviewees reported the frequency of meetings varied, from having contact six days a week to once every few weeks – far less than under YC. The differing amounts of support affected perceptions of how helpful control sessions were: '*[Practitioner] checks up, but if there's a problem, [practitioner] doesn't do much about it. So, basically, [practitioner] comes over for nothing or [practitioner] checks up for nothing.*' In Appendix B we present additional qualitative evidence from controls in relation to weaker ties being established between adolescents and practitioners, and session content being less based on CBT-techniques, but still having some focus on education and employment.[18]

3.5.3 Practitioners' Views on Your Choice

Practitioners had positive views of the training received to deliver YC, providing a clear sense of the practical application of CBT-tools. Even those with a background in psychology felt the training clarified how CBT-tools could be '*utilised most effectively.*' Practitioners also found it helpful to receive ideas for delivering sessions and creative strategies for engaging young people: '*I think a lot of the young people we work with, especially when reaching the threshold for Your Choice, are really in a stuck position. And being able to come up with actual interventions and ideas for how to move from that stuck position was really helpful.*' Mirroring themes from interviews with treated adolescents, Appendix B presents evidence on the views of practitioners, reinforcing the idea that YC helped build trusting adolescent-practitioner relationships, sessions used CBT-techniques, and hence the underlying theory of change was supported.

[18] Those treated youth with previous experience of statutory support services highlighted how activities were more interactive with YC, allowing them to engage in a wider range of conversations with their practitioner. Previous support was generally viewed as unhelpful or less helpful because standard sessions included '*just being talked to and you feel like you're punished.*'

Two issues were identified that signalled potential concerns for scale-up. First, in terms of buy-in, some practitioners were initially sceptical about YC, assuming it would not differ to what they already do or that it would be more like '*crisis management*' rather than supporting '*understanding of that distress.*' These doubts cleared up as YC was put into practice but challenges remained to convince them of the sustainability of the intervention. Second, concerns remained around workloads, leaving little capacity to complete and submit forms for data collection, for an already intensive intervention. Most practitioners mentioned needing to have their caseloads reduced to meet the time demands of YC: '*You have to be realistic about the case numbers that you have [...] And obviously you're asking for a commitment from the young person, so you need to make sure that you've got the calendar availability to honour that commitment that you're asking from them. Because often these are young people that have been let down repeatedly by their own, by the people around them as well as professionals, so its important.*'

Both issues were incorporated into the learnings from the pilot, but given the potential for each to lead to voltage drops, we come back to reconsider them in more detail using the qualitative data from the scale-up when discussing supply-side challenges to scaling more broadly.

3.5.4 YC Trainers and Clinical Leads

Clinical staff delivering the YC training to practitioners indicated it was helpful being given scope to adapt training sessions – for example, some found that four days was not long enough to deliver training and so they extended it. On the third ingredient of YC, clinical supervision sessions were seen as a necessary continuation of lessons learnt during training. This allowed practitioners to reflect about the impact of their work, but also the impact of the work on them, helping practitioners avoid experiencing trauma from working with vulnerable adolescents.

3.5.5 Allocators

Qualitative data was also collected to shed light on the fidelity of delivery with the experimental research design. We aimed to learn about how LAs allocated young people to teams via interviews with allocators, and a survey to operational leads ($n = 22$). Building on the evidence on balance (Tables 1 to 3), adherence to the research design was reinforced by the qualitative data, highlighting that allocation decisions primarily hinged on team capacity at the point of referral rather than individual characteristics or perceived gains from participation in YC. As one allocator described the assignment procedure: '*As new referrals came in, I would have a look at the reasons for the referral, and if they hit the sort of the Your Choice thing, I would give them to either unit one or unit two. So unit one was the control and unit two was the treatment team or delivery team, and then they would then try and get those young people signed up to the programme.*'

3.6 Learnings

Despite the challenges of introducing a new intervention for vulnerable youth across LAs embedded within adolescent services that already operate at near full capacity, the evidence shows some promise of the YC intervention. While no single piece of evidence is compelling, the combination of quantitative and qualitative evidence points in the same direction to suggest: (i) the fidelity of implementation, with the capacity-based assignment rule and the balanced recruitment conditions being supported; (ii) trusting relationships between adolescents and practitioners being built; (iii) the delivery of CBT-techniques in treated sessions; (iv) the primary outcome of the incidence of conduct problems being in the high/very high range being reduced; (v) most secondary outcomes moving in the direction predicted by the theory of change.

As a result of engagement with stakeholders and the co-design of structures in each LA to support the intervention and evaluation, the YC program had a high degree of buy-in from those delivering adolescent services. The challenges arising from the pilot mostly related to being able to conduct an evaluation at-scale. In the Appendix, we detail the mitigating strategies put into place to ensure the strongest possible evaluation at-scale. These relate to improved data collection practices to improve the recording of session forms data and reduce the time burden on practitioners, to reduce attrition, and recognizing the practicalities of session delivery – allowing practitioners to deliver 36 sessions to treated young people over 12 to 18 weeks (rather than within 12 weeks).[19]

4 Scale-up

During the scale-up, 1465 young people were enrolled across 122 teams in 30 LAs. 75 teams were assigned to treatment, with 854 (58%) adolescents being assigned to treated teams and 611 assigned to controls – very similar proportions as in the pilot. Panel B of Table A1 shows the most teams participating in the scale-up are in *Adolescent Safeguarding*. This composition differs from the pilot in part because of the need for LAs to provide a fresh list of teams available to be trained. We later return to consider how changes in service composition at scale affect the interpretation of results. As detailed in the Appendix, the scale-up remained powered to detect if the intervention lived up to the promise shown in the pilot.

4.1 Balance

Table 5 shows baseline characteristics of adolescents in the scale-up. For each, we report the mean in treatment and control. As treatment is randomized within LAs, each providing a different combination of teams and serving vulnerable youth with slightly different needs, compositional differences may show up in the direct comparison of treatment and control means. In Column 3

[19] During the pilot, those aged 11-17 at the time of recruitment were considered eligible for YC. For scale-up, the age range was extended to include those who would turn 18 during the intervention.

we therefore show differences net of LA fixed effects (the randomization strata), and Column 4 provides the corresponding p-value on a t-test of the equality of conditional means.

Counter to earlier at-scale evidence on CBT-interventions to vulnerable youth [Blattman *et al.* 2017, Heller *et al.* 2017], around a third of our sample are female. There is an even split of younger and older adolescents, 84% of controls are from minority backgrounds, and they have high rates of special educational needs (12%), disability (20%) and being in social care (13%). Among controls, their most common primary need relates to socially unacceptable behavior (38%). The majority report being in school/college (83% in controls). The within LA differences in Column 3 are all relatively small, and the sample is balanced across all but two characteristics.

Table 6 shows baseline measures of our primary and secondary outcomes. 25% of controls have an SDQ score in the high/very high range for conduct problems. Around 30% of controls score in the high/very high range for hyperactivity, 19% for emotional problems, 30% for peer problems, and 18% score in the low/very low range for prosocial behaviors. The final rows report the other secondary outcomes at baseline: the social connectedness and prosocial identity scales.

Column 3 shows the within-LA differences are smaller than the raw differences. Within randomization strata, the samples are balanced on the five indicators for whether the adolescent lies in the extreme range although there are differences at the 5% level in the continuous score measures for prosocial behaviors ($p = .042$) and the prosocial identity scale ($p = .042$) – where treated adolescents score lower than controls on both. Table A2 shows balance on four out of five subcomponents of the social connectedness scale.

Table 7 examines assessed risks. Panel A shows risk assessments by young people are balanced on five out of six margins – the exception being that while 9.4% of treated children report that in the next month they are likely/very likely to get into trouble with the police, this is 3.5% among controls ($p = .045$). As with the pilot, Panel B shows that practitioners generally assess young persons to be at higher risk than the individuals themselves assess. Practitioners express higher concerns in relation to treated young people for nearly every risk type (except them being a witness to crime), and these differences remain within randomization strata.

Two points are of note. First, the results from Tables 5 to 7 suggest the samples are balanced on many characteristics of adolescents, primary and secondary outcomes, and young persons' self-assessment of risk. There is a more striking imbalance in practitioner assessments of risks young persons are exposed to. While this remains consistent with randomization protocols being followed at an aggregate level, it highlights that either: (i) the balanced recruitment condition might have been violated; (ii) practitioners tasked with delivering the YC intervention view young people differently in the scale-up, when the delivery of YC becomes more established.

Second, contrasting the pilot and scale-up samples, the scale-up sample better captures the regular London-wide flow of young people at medium/high risk of exposure to violence, while the pilot also reflects similarly assessed youth in LAs and services able to engage with the pilot phase. Those in the scale-up remain highly vulnerable – they are two to three times more likely to lie in

the extreme range of the five SDQ-domains than does the UK-wide population of 11-17 year olds. At the same time, the pilot sample of young people present worse characteristics at baseline, in terms of special education needs, disability, SDQ scores, and risk assessments. This is an issue we return to later to understand whether sample unrepresentativeness at scale-up represents a challenge to scaling and helps explain a loss in voltage.

4.2 Attrition

Of the 1465 adolescents recruited during the scale-up, 1066 were tracked to endline, an 27% attrition rate, far lower than for the pilot (50%). At scale-up, attrition is higher among the treated (31% vs. 20%) while the opposite was true in the pilot (45% vs. 57%). The attrition rate is lower than assumed in the power calculations for the scale-up, reflecting the mitigation strategies put into place. The second part of Table A3 examines correlates of attrition at scale-up and shows: (i) unconditionally, attrition is not statistically different between treatment and controls (Column 4); (ii) this remains the case conditional on LA fixed effects (Column 5), and adolescent characteristics (Column 6); (iii) unlike for the pilot, there is no evidence of heterogeneous attrition on observables: the joint F-test of interactions of characteristics with treatment is not significant ($F = .783$). This evidence all moves in the same direction of suggesting any voltage loss is not due to the level or nature of selective attrition.

4.3 Implementation

We draw on two sources of data to understand YC delivery at-scale: data from young person-practitioner sessions, and endline data to young persons and practitioners.

On session data, 80% (57%) of treated (control) practitioners submitted at least one completed session form – far higher than in the pilot. In total, 5860 (1711) session forms were collected in treated (control) groups. Panel A of Figure 3 shows a histogram of the number of young person-practitioner sessions received by treatment and control adolescents: (i) far more sessions took place under YC than for controls; (ii) 27% of treated adolescents received zero sessions, while this rises to 48% for controls; (iii) the average number of sessions (conditional on at least one session) is 9.5 (5.3) for treated (controls), so that even during the scale-up with the extended delivery timeline, only 30% of treated adolescents received 10 or more sessions (including zeroes), with only 2% of treated youth recorded to have received 30 or more sessions.

Panel B shows session content, linking to the use of YC materials and CBT-tools: (i) 96% of recorded sessions for the treated group had YC content; (ii) 59% of treated sessions record the use of CBT-tools, 33% record relationship building as a part of their content; (iii) although far fewer sessions take place in controls, among those that did, around 20% of them report using materials from the YC curriculum, while only 13% record the use of CBT-tools and only 10% record a

focus on relationship building. While this might reflect some leakage of YC into controls, it can also reflect the fact that, as highlighted in the qualitative evidence from the pilot, practitioners can sometimes use CBT-style tools or emphasize relationship building in their business-as-usual provision of statutory services to vulnerable youth.

In summary, the sessions data suggests implementation during the scale-up remained far from perfect, especially in terms of the intensity of treatment relative to what was designed. Although it is hard to draw firm conclusions given the patchiness of session forms data, the evidence suggests the likelihood of no session taking place fell for treated youth (27% at scale-up from 37% in the pilot) and controls (48% vs. 65%), and conditional on at least one session with a young person, almost the same number of sessions were delivered in treatment (9.5 at scale-up vs. 8.9 in the pilot) and control groups (5.3 vs 5.0). Hence key ingredients of YC continued to be delivered at-scale and the fact that more sessions were delivered at-scale than in the pilot further suggests that any reduced intensity of treatment – or a loss in compliance voltage [List 2026] – is unlikely to be driving voltage drops.[20]

A second source of data we can use to additionally explore the issue is through questions asked to young people and practitioners at endline designed to measure the fidelity of implementation. These questions were designed to measure the extent to which practitioners were delivering content in line with YC training. Young people and practitioners were asked seven questions related to sessions content: Table C3 details the questions used. We use these to construct overall fidelity scores, and then separately standardize the scores from young people and practitioners. Using these scores as an outcome, we then regress them on a treatment indicator. The results are in Table A9, where Columns 1 and 2 show results from the pilot, and Columns 3 and 4 show results from the scale-up (where we additionally control for local authority fixed effects). The treatment effect on fidelity scores are insignificant in the pilot, but are larger and significant in the scale-up when questions are asked to young people ($p = .050$). In line with the session forms data, this continues to suggest key ingredients of YC were delivered at-scale.

4.4 Results

To estimate treatment effects of being assigned to a YC team relative to a control team, we use the following specification for young person i in local authority l:

$$y_{il} = \alpha + \delta y_{0il} + \beta T_{il} + \lambda_l + u_{il}, \tag{2}$$

where y_{il} is the outcome of interest, all controls are as previously defined in (1) and λ_l are local authority fixed effects (the randomization strata). The β estimate is interpreted as an ITT parameter, and we again follow Chaisemartin and Ramirez-Cuellar [2024] to cluster standard errors

[20] We explore the issue further in the Appendix, using an IV-approach to estimate dosage effects of YC.

at the local authority level. We report conventional p-values on the null that $\beta = 0$, and where appropriate, we report p-values that adjust for multiple hypothesis testing using: (i) the stepdown method [Romano and Wolf 2005, Romano *et al.* 2008, Romano and Wolf 2016]; (ii) the List *et al.* [2023] procedure given our covariate-adjusted, fixed-effects specification.

Table 8 shows treatment effect estimates for our primary and secondary outcomes, where the SDQ-subscale measures all relate to whether the adolescent lies in the extreme range of each. On the primary outcome, Column 1 shows the likelihood of high/very high conduct problems *rises* slightly by 2.8pp (from a baseline mean of 25% in controls tracked to endline). While this is not statistically significant ($p = .328$), the 95% confidence interval rules out any fall larger than 2.9pp corresponding to 11% of the baseline mean in controls – so far lower than the point estimate reduction observed in the pilot. At the foot of each column, we report an F-test of equality of treatment effects between the pilot and scale-up: this shows that for our primary outcome, the impacts differ for the pilot and scale-up ($p = .000$).

On secondary outcomes, the point estimate impacts on the likelihood of high/very high problems related to hyperactivity and emotional problems are both negative, and both differ from those in the pilot: the impact on hyperactivity is less negative ($p = .019$), while the impact on emotional problems is more negative ($p = .011$). Column 4 shows the likelihood of being in the high/very high range for peer problems falls by 3.2pp (corresponding to 11% of the baseline mean in controls), and rises for prosocial behaviors by 2.5pp (14% of the baseline mean in controls). Neither of these estimates are statistically significant, nor does either differ from the pilot estimates. The remaining Columns in Table 8 show treatment effects on other secondary outcomes. The point estimates on outcomes related to mechanisms in the theory of change such as social connectedness and emotional self-regulation move in the desired direction, as do some of the intended outcomes such as the self-assessed risk assessment, but none of these impacts are statistically significant. Nor do they differ from the estimates from the pilot.[21]

Table A16 uses IPWs to adjust for selective attrition (even though at scale-up, there was no evidence of heterogeneous attrition on observables). Treatment effects on our primary outcome remain small, positive and not statistically different from zero, but a few secondary outcomes are marginally significant. For example, the social connectedness score increases by .085 ($p = .074$ and this is robust to adjusting p-values for multiple hypothesis testing ($p = .066$).

[21] This pattern of findings continues to hold if: (i) a conditional logit specification for being in the extreme range of each SDQ-subscale is used (Table A10); (ii) we consider the continuous scales for each SDQ measure (Table A11); (iii) we additionally control for characteristics imbalanced at baseline (Table A12); (iv) we use post-double LASSO to select for the baseline covariates to control for (Table A13); (v) we cluster standard errors at the team level (Table A14); (iv) we use bootstrapped standard errors that are clustered at the team level (Table A15).

4.5 Qualitative Evidence

Despite the promise shown in the pilot, the YC intervention failed to be effective at-scale for the primary outcome and most secondary outcomes. A natural explanation might be a loss of fidelity with the research design at-scale. Although the earlier evidence suggested no measurable worsening in adherence to random assignment, we use the qualitative data to shed more light on the issue.

We find that the qualitative evidence from young persons and practitioners during the scale-up reaffirms what was previously documented during the pilot. Specifically: (i) young people viewed relationship building as key to the success of YC, enabling them to talk openly to practitioners; (ii) young people and practitioners reported the use of CBT-tools in YC sessions (consistent with sessions form data presented in Figure 3); (iii) the underlying theory of change related to mechanisms and intended outcomes continued to be supported from both perspectives, with evidence of changes being sustained over time for young people. Given these perceptions largely mirror those from the pilot, we leave their description to Appendix B.

This all reiterates that although the implementation of YC was far from perfect and less intensely delivered than designed, the intervention was implemented similarly in pilot and scale-up phases, and the perceived benefits of the intervention remained highly visible at-scale to practitioners and treated adolescents. This helps rule out two common explanations for voltage drops: worsening adherence to fidelity, or a loss in quality voltage [List 2026] driven by a lack of engagement with the intervention by practitioners and young people.

In the remainder of the analysis, we drill down to better understand the key challenges that prevented Your Choice from scaling.

5 Identifying the Challenges to Scaling

Many interventions that show promise in small-scale pilots fail to deliver results at-scale [List 2022, 2024], including for criminal justice interventions [Stevenson 2023]. Beyond concerns over a loss in fidelity at-scale, we follow the typology described in List [2026] to examine four broad challenges to scaling: (i) sample unrepresentativeness, where participants at-scale differ in characteristics or motivation from those involved in the pilot; (ii) situational unrepresentativeness, where conditions related to the intervention differ at-scale from those in the pilot; (iii) supply-side constraints, where organizational uncertainties, complexities and spillovers only emerge at-scale; (iv) increasing marginal costs of delivery.

We use the qualitative and quantitative data to understand whether these concerns represent real challenges to scaling the YC intervention and thus lie behind voltage drops. At the end of this Section (and Table 12), we summarize our findings and where for each potential challenge, we highlight where the qualitative and quantitative evidence aligns or not.

5.1 Sample Unrepresentativeness

A common cause of voltage drops is sample unrepresentativeness, whereby initial pilot participants differ from those recruited at scale-up. Changes in sample characteristics and motivation can arise from: (i) adolescent services in LAs that engage with the evaluation at-scale differing from those that participate in the initial pilot; (ii) characteristics of adolescents differing between pilot and scale-up because the latter phase captures the regular flow of young people assessed to be at medium/high risk of exposure to violence, while the pilot sample better reflects similarly assessed individuals in LAs and services able to engage with the initial implementation of YC; (iii) utilizing less motivated or supported practitioners at-scale. We examine these in turn.

5.1.1 Services

We first consider whether the change in composition of services that young people are assigned to between the pilot and scale-up explains the loss in voltage. Table A17 shows the results in the scale-up when we additionally control for service fixed effects in specification (2). The estimates remain mostly in line with the earlier results. For example, we find the likelihood of being in the high/very high range of conduct problems rises by 4.2pp, the 95% confidence interval rules out any fall larger than 1.7pp corresponding to 7% of the baseline mean in controls. This remains far from the finding in the pilot that Your Choice led to an 11pp fall in the likelihood of being in the high/very high range of conduct problems. All other secondary outcomes also remain of the same sign as in our baseline results at-scale, and with similar magnitudes.

5.1.2 Adolescents

The evidence on balance at scale-up shows that characteristics of young people within LAs – in terms of demographics, primary need, activities, being in the extreme range on each SDQ-subscale, or self-assessed risks – remain similar between treatment and controls. This reinforces the idea that any voltage drop does not originate from a reduction in fidelity with the research design. Rather, the pilot sample presents worse characteristics at baseline relative to those recruited during scale-up: they are more likely to have special education needs or a disability, more likely to be in the extreme range of conduct problems, emotional problems, and self-assess to have greater risk of exposure to violence. This change in composition of vulnerable youth can lie behind the voltage drop – especially when the pilot results already hint at the intervention having more pronounced effects on those with the most severe conduct problems to begin with, consistent with existing evidence [Landenberger and Lipsey 2005, Koehler *et al.* 2013, Bhatt *et al.* 2021].[22]

[22] The potential for there being heterogeneous effects of YC depending on adolescent characteristics was highlighted in the qualitative evidence from the pilot. Practitioners reported that young people who understand themselves as being a victim of coercion would likely be more receptive. They also mentioned that young people may need a certain level of maturity or emotional functioning to understand the intervention's concepts and suggested that older children might be better suited to YC. There was also some mention that those with more experience of

To assess how the change in composition of young people from the pilot and the scale-up might drive the voltage drop, we proceed in three steps.

First, in Tables A18 to A20 we report for each characteristic considered in the balance tables, the overall means in the pilot and scale-up trials, their (normalized) difference, and the associated p-value of a test of equality of means. On background characteristics, those recruited at scale-up are similar to those recruited into the pilot on many dimensions, but they are more likely to be from a minority, less likely to report a disability, and have slightly different primary needs (Table A18). In terms of the SDQ measures, scores for those recruited into the pilot are systematically worse than for those recruited at scale-up, even though there are no statistically significant differences between the samples (Table A19). Risk assessment is where we see the most pronounced changes between those recruited at the pilot and scale-up: those recruited at scale-up self-report being exposed to significantly fewer risks (1.8 vs. 2.0, $p = .018$) and practitioners also report young persons recruited at scale-up to be exposed to significantly fewer risks than those recruited at the pilot (2.6 vs. 3.0, $p = .000$) (Table A20).

Second, we follow the entropy balancing approach of Hainmueller [2012] to reweight the scale-up sample to match baseline means among the pilot sample. Table 9 presents the results, focusing on our primary outcome of scoring in the high/very high range of the conduct problems SDQ-subscale. As a benchmark, Columns 1 and 2 show the results from the pilot and scale-up. In Column 3 we reweight for differences in service assignments and baseline characteristics of youth. We find that at scale-up, the point estimate of the treatment effect of YC becomes negative and much more in line with the pilot results. Specifically, the likelihood of being in the high/very high range of conduct problems falls by 16pp, corresponding to 67% of the baseline mean in controls. In Column 4 we repeat the approach but use a LASSO regression to first select those characteristics of young people differing between the pilot and scale-up. When we reweight the scale-up estimates we then almost perfectly replicate the estimated treatment effect point estimate in the pilot for our primary outcome: the likelihood of being in the high/very high range of conduct problems falls by 10.6pp, corresponding to 45% of the baseline mean in controls.[23]

Table A21 repeats the analysis for all other outcomes, using the LASSO specification from Column 4 in Table 9. Panel A shows the pilot results as a benchmark. Panel B shows that for outcomes related to being in the extreme range of peer problems, the social connectedness scale, the wellbeing index and self-assessed risks, the reweighted treatment effects at-scale move towards those found in the pilot. This all supports the theory of change holding at scale.

To summarize, the loss in voltage on the primary outcome, and some of the secondary outcomes,

victimization, rather than perpetration, had better outcomes as a result of the intervention.

[23] Since observations with very high weights could dominate estimates based on entropy balancing, we also re-estimate the model after trimming the top 5% of the distribution of weights. Our findings are not driven by a small number of observations: when controlling for all covariates, the results suggest YC reduced the likelihood of being in the high/very high range of conduct problems by 18.1pp; the corresponding figures for the LASSO selected covariates is 14.4pp.

can nearly all be explained by change in composition of young people recruited into the scale-up relative to the pilot. The scaled up sample has a lower representation of young people with severe difficulties and high risk (from the point of view of the young person). An implication of this is that intervention impacts are heterogeneous, and that the pilot sample had a better representation of the types of young people who benefit more from the intervention.

In a third step, we therefore investigate whether impacts at-scale do indeed vary by baseline SDQ scores, other secondary outcomes and risk assessments. For each outcome we estimate whether there are heterogeneous treatment effects for those in the problematic range for the same outcome at baseline. For example, for SDQ score outcomes, we examine heterogeneous effects by whether the individual lies in the high/very high range on that domain at baseline. For non-SDQ outcomes we use a 30% threshold to define the problematic group at baseline to match the shares for the SDQ measures.

The results are in Table 10. For the primary outcome, Column 1 shows a split by whether the young person has high/very high conduct problems at baseline. We find the intervention reduces the likelihood of being in the high/very high range significantly more for those that were in this range to begin with: the differential impact across groups is 13.6pp ($p = .059$). Similarly, Column 2 shows the intervention reduces the likelihood of being in the high/very high range of hyperactivity by significantly more for those that were in this range at baseline: the differential impact is 21.1pp ($p = .005$). We also find heterogenous responses along two other margins of secondary outcomes: the treatment effect on the social connectedness scale is significantly greater for those in the problematic range of this outcome at baseline (25.1pp, $p = .018$), and finally, self-assessed risk scores fall by significantly more for those who faced greater self-assessed risks at baseline ($p = .077$). Overall, the evidence points to the intervention having greater impact on the the primary outcome and some secondary outcomes for those more vulnerable to begin with, a group that were recruited into the pilot to a great extent than those recruited at scale-up.

5.1.3 Practitioners

The third source of sample unrepresentativeness we examine originates from practitioners. In our context, as the intervention is delivered within pre-existing constraints of adolescent services, including utilizing the existing practitioner workforce, the issue is not caused by an inelastic supply of high-quality practitioners as the intervention scales [Davis *et al.* 2017]. Rather the potential issue is changes in motivation, training or supervision of existing practitioners at-scale.

On motivation, we compare how practitioner's risk-assessment of young people differs between the pilot and scale-up (Panel B in Tables 3 and 7). In line with the changed composition of young people at-scale, practitioner's risk assessments of young people recruited at-scale are lower than those recruited in the pilot. At the same time, within the scale-up phase, the treatment-control difference in practitioners' assessment of the risks young people face is more pronounced. This

can reflect that as aims and delivery of YC become more established, practitioners delivering YC view young people differently from the outset of the intervention. This kind of dynamic might then naturally lead to a voltage *increase* if practitioners try harder to assist young people given their increased perception of multiple risks the individual faces.

On training, the idea that practitioners are less well trained in the scale-up than the pilot is not supported by the qualitative evidence. This shows a strict adherence to training protocols, with a large fraction of practitioners assigned to YC undergoing training, and the vast majority of practitioners in control teams not undergoing any training.

The final margin related to practitioners that might lead to a voltage drop is a weakening of the clinical supervision that practitioners received during scale-up – a core ingredient of YC. However, this idea is also not supported by the evidence. Regular clinical supervision of practitioners continued during the scale-up: 1593 YC supervision sessions were delivered to 570 treated practitioners, and 135 supervision sessions were delivered to 234 control practitioners.[24]

Overall, the evidence does not suggest that practitioner-side issues worsened during scale-up and contributed to sample unrepresentativeness, hence are not a source of voltage drops.

5.2 Situational Unrepresentativeness

A second class of challenges to scaling we consider are issues around situational unrepresentativeness, where conditions related to the intervention differ at-scale from those in the potentially more controlled environment of a pilot. We examine some of these using the qualitative data from scale-up that involves interviews with: (i) young people assigned to YC ($n = 22$ across 13 LAs); (ii) practitioners delivering YC ($n = 20$ across 16 LAs); (iii) YC trainers ($n = 9$ across six LAs); (iv) other leads from LAs ($n = 12$).[25]

In line with the design of YC, the flexibility and adaptation of the intervention continued to be seen as central to its success. This shows up in numerous ways in the qualitative evidence. For example, LA leads referenced the importance of being responsive to young people's needs in the frequency of contact with YC practitioners: '*If they don't want to meet three times a week, then we won't push it, because we run the risk of them dropping out.*'; '*If the young person's only happy to engage once a week or if you can keep in contact with them through texts and email or WhatsApp, then you know, that actually is good enough. Because it is about building the relationship.*' Practitioners also described tailoring YC to meet young people's needs, making adjustments to account for learning difficulties, speech/language needs, or neurodivergence.

[24] In line with experiences during the pilot, practitioners in YC did not receive as much supervision as intended – supervision data suggests the share of practitioners who received three or more supervision sessions in treatment (control) was 28% (10%) – but this lack of intensity does not differ across study phases.

[25] On (i), we interviewed 22 young people about their experiences and perceptions of intervention impacts. These interviews were conducted around 15 weeks after their recruitment. We re-interviewed 12 of these young people six months later, and so can also shed light on slightly more persistent changes in perception than we were able to do in the pilot. Interviewees were on average aged 15, with a near equal split by gender.

Although the flexibility and adaptability of interventions is often seen as a key to scaling, our evidence reveals two costs of building in such flexibility. The first is that the willingness to tailor delivery to specific needs of adolescents meant there was a shortfall in the number of sessions that young people received relative to the intended intensity of treatment. This potentially reduces the efficacy of the intervention relative to the intended design, although as evidenced earlier, the number of sessions increased from pilot to scale-up – so this shortfall occurred over both study phases and therefore is not key to understanding the voltage drop observed. Second, flexibility in delivery is a double-edged sword in that there is no single formula or recipe to be passed on to ensure successful delivery to adolescents. This is especially problematic for an intervention operating within real world constraints of adolescent services that rely on: (i) the adoption of a train-the-trainer model where clinical leads train the next wave of practitioners in YC; (ii) an emphasis on clinical supervision enabling practitioners to learn from others' experiences. This cause of voltage drops echoes one of the reasons given for the failure to scale of the *Becoming a Man* intervention in Chicago [Bhatt *et al.* 2021].

5.3 Supply-side Constraints

A third class of challenges to scaling we consider are supply-side constraints, where organizational uncertainties and complexities only emerge at-scale. These potential causes of voltage drops relate to system-wide factors, and many are hard to measure quantitatively. Our qualitative data sheds light on three potential sources of supply-side constraint.

5.3.1 Resistance and Buy-in

A natural tension arising when interventions scale is that new programs viewed as most different from business-as-usual can face more resistance, despite being potentially effective. This is a kind of stabilizing force that often prevents criminal justice interventions from scaling [Stevenson 2023], has parallels to the much studied non-adherence problem in medicine [Al-Ubaydli *et al.* 2017], and can be driven by individuals experiencing psychological switching costs when modifying behavior or a desire to conform to established workplace norms [Klemperer 1987, Carroll *et al.* 2000]. As with the pilot, resistance to change during scale-up was reflected in qualitative interviews with LA leads: '*I think, initially, there was a lot of persuading, convincing that needed to occur. It was new. It was a change.*'

A related issue that can slow down the adoption of new practices arises from the fact that those working in adolescent services are subject to a constant flow of innovations. Each one is perceived to come and go, offering staff little incentive to make intervention-specific investments. Ironically, this dynamic might slow down the adoption of new practices that genuinely lead to better outcomes in organizations that are most willing to experiment with new ways of working. This concern was raised by YC trainers, with some practitioners' perceiving YC as '*just another*

programme that's going to come and go,' and contributing to their reluctance to engage.[26]

As described earlier, our study represents a collaboration with London-wide stakeholders and top-level administrators in each LA. Their engagement helps signal to practitioners the political will and buy-in of senior leadership to drive change. Moreover, the coordinated introduction of the intervention across LAs can help break the inertia of established workplace practices, and shift practices towards a new norm [Chassang 2010]. Moreover, a key feature of the intervention that helped reduce resistance and gain practitioner buy-in was that YC was perceived to add to practitioners' toolkit, providing opportunities for their professional development. Another theme that came across from practitioners was their sense that YC brought needed structure to their work and the young people they worked with: '*This actually feels really monumental, the fact that like we have a regular structured thing in his life.*'[27]

YC trainers also felt training had helped practitioners develop the skills needed to support the complex needs of young people: '*They will manage extremely high-risk cases, children with multiple complex needs. But because they have got their skills upskilled, they are better able to engage with families for providing resources at an early age and at an early stage before they escalate to needing, probably, a high-intensive, high-cost placement.*' The transferability of new skills was also valued – practitioners felt they would continue using YC CBT-techniques: '*Just having access to all of those is really useful and I would definitely keep using them.*'

These elements of top-down buy-in combined with the demonstration of value-added to practitioners, helped increased bottom-up buy-in over time. As such, these dynamics would likely increase voltage at scale-up, all else equal.

5.3.2 Sustainability

Some uncertainties are only realized at-scale. We consider two sources of uncertainty related to the perceived sustainability of the intervention.

First, a key constraint only realized at scale-up is staff turnover. This is relevant in our context: childrens' services across LAs operate at near full capacity and with fixed resources, despite increasing demands on them given long run trends of rising youth exposure to violence. The issue of staffing was highlighted by SPoCs as negatively impacting services' capacity to implement YC: '*Your Choice got, kind of, parked because our general bread and butter wasn't even being done,*

[26] This relates to the concept of stabilizers leading to voltage drops, or socioeconomic forces that resist externally-imposed change [Stevenson 2023]. This is relevant in our study, in which target population has a complex set of needs. Without sustained effort by practitioners or the intervention remaining in place, all else equal, these contextual factors mean it is harder for a beneficial cycle of change that accumulates over time to be enacted, and far easier for participants to drift back to their initial circumstances.

[27] YC was also seen by SPoCs as bringing structure to practitioners' work with young people through its intended three-session per week framework, its provision of activities, and its use of goal setting: '*The demand that children get seen three times a week and then, in a sense, that forced our staff to actually make timetables for the child and themselves. You know, that, for me, gave them much more structure in terms of how they were working with the child and their family.*'

because we lost so many people.' Staff turnover meant that young people's involvement in YC could end if their practitioner left, or services could lose entire practitioner teams. Turnover of clinical leads in LAs led to particularly acute strains on voltage because of the consequent reductions in clinical supervision, and a reduced ability to train incoming practitioners.

The inability to smooth shocks to staff turnover relates to the wider issue of uncertainty over future funding for the intervention. Pilot phases often have the carrot of future funding becoming available to scale-up promising interventions. While it is well appreciated that such sequencing can distort incentives, leading to false positives from pilots and bandwagon effects [List 2024], when interventions are rolled out at-scale implementers usually face greater uncertainty over future funding and other elements supporting sustained change – and this uncertainty can lead to voltage drops. Our qualitative evidence sheds light on this concern.

On the one hand, practitioners felt they would continue using elements of YC, such as CBT-techniques, goal-setting and activities: '*I will definitely be using some of the stuff that I learned with the young people wherever it fits.*' SPoCs and LA leads also mentioned resources and CBT techniques provided through YC could continue to be used beyond the funded life of the intervention: '*I do think people that have used it, have seen the value in using the tools. So, I think they would continue to use the tools going forward.*' However, they also felt a lack of funding would affect future implementation because of a lack of ability to support activities for young people, and the ability to retain clinical leads: '*Without the funding I don't think it will carry on.*' Accessing funding was also a key concern expressed by trainers: '*It would completely come down to whether the money is there or not.*' Ultimately, decisions about the continued delivery of YC were felt to rest with management and funding availability: '*It's got to be very specific buy-in from managers across the different services. And then somehow support staff to have time do it. Because it is a little bit extra work, it's not nothing.*'

These kinds of concerns over sustainability only gradually build up during scale-up. We suggest this can apply more widely in public services that already operate at full capacity pre-intervention. We come back to the issue when considering changes in the costs of delivering YC between the pilot and scale-up stages.

5.3.3 System-wide Constraints

Another complexity that only emerges at-scale is binding system-wide constraints that lead to voltage drops. For example, practitioners felt that if the intensive work around YC was to continue, they would need to have workloads reduced. Despite mitigation strategies to make data collection in sessions forms easier for practitioners, operating at scale increased the demands on capacity as practitioners were required to recruit and deliver the intervention to more young people than in the pilot. Practitioners consistently reported the challenge of delivering the intense level of contact for a young person receiving YC while maintaining their usual case load. The additional

work that YC had presented meant they still felt they had competing priorities: '*If I knew that my job was solely to work with the young person, this would work much smoother. It would work smoother. But my job is not that. My job is whole family. So that dilutes my attention.*' Trainers also referenced practitioners' existing workloads as reducing their capacity to engage in with YC: '*Whenever you try to get practitioners to do something new or something that requires another form to be filled in, you're always going to have resistance. That's because people are doing the best that they can, they're pushed to their limit. So, straight away it never... I don't think whatever you introduce, it never goes down very well.*' Finally, this issues links back to earlier concerns related to resistance to change: SPoCs and LA leads described that where YC presented a departure from usual ways of working, it was a harder sell to practitioners to add to their workloads.

5.3.4 Spillovers

This leads naturally to concern over general equilibrium effects or spillovers that are only realized at-scale, leading to voltage drops. In our study context, binding capacity constraints can mean any additional attention, effort or time devoted to treated adolescents has detrimental spillovers either onto adolescents assigned to control teams or to those assigned to teams not in our study sample. If these kinds of crowding out are noted by implementing staff, they can reduce buy-in for the new intervention. This concern is partly ameliorated by the way that teams and practitioners are organized. Specifically, practitioners work within the same team – they do not typically divide their time between teams. Similarly, all youth assigned to a team have the same treatment status – so there is no margin for treated practitioners to reallocate effort between control and treated adolescents. Where unintended spillovers could arise is from the fact that leads in LAs – either the SPoCs, Clinical Leads or YC trainers – might reallocate effort away from vulnerable adolescents either in control or non study teams towards those in treated teams. We cannot measure this type of top-down reallocation of effort in our quantitative data, but the fact that childrens' services operate at capacity suggest that without efficiency gains or additional resources, there can be unmeasured opportunity costs of the intervention through such crowd out.

5.4 Costs

A final explanation for voltage drops relates to cost inflation, or the average cost of delivering the intervention rising at-scale [Davis *et al.* 2017, Bhatt *et al.* 2021]. We can tentatively examine the issue using cost data from the pilot and scale-up – subject to the caveat that only a handful of LAs were able to provide sufficiently detailed costings across 19 line items split between fixed and recurring costs. These measured costs represent the total cost of delivering YC, not the cost net of providing business-as-usual services to controls. For each study phase, we match these costs to the

estimated flow of adolescents that YC was delivered to by a standard sized team of practitioners.[28]

The cost data is presented in Table 11, split into: (i) set-up costs (Panel A); (ii) recurring costs (Panel B). For each line item, we show costs during the pilot (Columns 1 and 2), and during scale-up (Columns 3 and 4). To make like-for-like cost comparisons, the first row shows the average flow of participants into YC teams in a month – this rises from 2.7 in the pilot to 3.8 at-scale. Columns 2 and 4 show the per-participant costs for each line item at the pilot and scale-up phases, and Column 5 shows the ratio of per participant costs, to identify which elements drive changes in delivery costs at scale.

In the pilot, the average per participant cost of delivering YC by a practitioner team was just over £9,000 (Column 2, row 22): 53% of these related to set-up costs. At scale-up, the average per participant cost nearly doubled to over £17,000, tentatively suggesting the emergence of diseconomies of scale being another potential cause of voltage drops. Indeed, the magnitude of this change is very much in line with an emerging general lesson from the science of scaling, that cost inflation at-scale can lead to costs to be near double those in pilots [List 2026].[29]

At-scale, the balance of costs shifts away from set-up costs, that represent 36% of total costs, towards recurring costs. To dig deeper, we examine specific line items to understand how the cost structure changed at-scale and identify how the challenges to scale described earlier on situational unrepresentativeness and supply-side constraints then translate into rising costs. We do so by considering the ratio of per participant costs for each line item shown in Column 5 of Table 11. Three points are of note.

First, average fixed costs per participant increase at-scale by 27% (row 13): the items that cause this are increased costs relate to data collection and governance (row 8), and the participation of clinical leads, YC trainers and practitioners into training (rows 6 and 10). This might reflect the challenge raised earlier, that staff turnover among senior leads in LAs prevents gains from specialization to be fully realized at-scale, and that local authorities had to move up the labor supply curve as they seek to recruit given staff turnover [Davis *et al.* 2017]. Second, average recurring costs per participant more than double at scale-up (row 21). The line items causing this relate to: (i) clinical supervision (rows 14 and 16), that again might reflect consequences of turnover of senior staff; (ii) 'step down' costs (row 20) – that are costs incurred as local authorities provided resources to enable young people to continue with elements of YC to sustain its impact.

[28] Diseconomies of scale can arise from either increases in participation or implementation costs. In our context only the latter are relevant – as the flow of young people assessed to be at medium/high risk of exposure to violence is taken as given by each local authority.

[29] As a point of comparison, we can compare these with estimates of the unit costs of violent crime. Heeks *et al.* [2018] provide the best estimates of these for the UK, taking account of: (i) anticipation costs (e.g. defensive investments); (ii) consequences of crime (e.g. value of property stolen or damaged); (iii) costs in response to crime for the police and criminal justice system. Their estimates suggest the social cost of being a victim of a violent crime without injury is £5,930. Hence in a best case scenario, the YC intervention would be plausibly cost effective at-scale if it reduced the number of violent incidents without injury by three for each treated young person. Around 10% of controls in the scale-up reported they would be likely/very likely to physically hurt someone in the month after the baseline (Table 7).

That is in line with the qualitative evidence on sustainability that suggested practitioners remained keen to support young people through YC. Finally, given fixed aggregate budgets for childrens' services as a whole, the cost increases associated with YC might then lead to detrimental spillovers through a crowding out of services to non-YC related services.

5.5 Summarizing the Challenges to Scaling

We bring together the evidence to summarize challenges to scale in Table 12, following the frameworks described in List [2024, 2026]. For each, we describe whether each piece of evidence suggests it is a challenge to scaling in our study context, and whether the quantitative and qualitative evidence aligns. The first challenge relates to the generation of false positives from pilots. Here we draw together the quantitative and qualitative evidence, that all aligns, to present multiple indications why YC had the potential to reduce young person's exposure to violence in a way set out in the theory of change. Rows two to five then summarize the evidence on the core challenges we have evidenced: sample unrepresentativeness, situational unrepresentativeness, supply-side constraints, and cost inflation. The last challenge highlighted – on general equilibrium or spillover effects is more conjectural given our data.

6 Discussion

The *Your Choice* intervention showed promise at a pilot stage but faced challenges to scale. This is despite having top-down buy in from London-wide stakeholders from the outset, the intervention being designed by these stakeholders, the research design recognizing feasibility constraints in how statutory services are delivered to vulnerable adolescents in LAs, and all materials being co-designed with the needs of the targeted group in mind. We use our evidence to set out implications for option-C testing in similar contexts – so how to go beyond A/B testing and build into research designs treatment arms that shed quantitative light on what might ameliorate voltage drops.

The challenges to scaling we identify through the combination of quantitative and qualitative evidence raises multiple issues for option-C thinking around causes of voltage drops.

First, our quantitative evidence highlights the relevance of sample unrepresentativeness at-scale. Those recruited at-scale are less vulnerable than those in the pilot, and in line with earlier literature, the evidence suggests YC might have more pronounced impacts on those more vulnerable to begin with [Landenberger and Lipsey 2005, Lipsey *et al.* 2007, Koehler *et al.* 2013]. While the intervention might have been implicitly designed with the most vulnerable youth in mind, the basic fact is that the natural inflow through referrals of such highly vulnerable youth is not sufficiently high, even in a London-wide experiment. Such issues might need to be borne in mind whenever the eligible population – even though all are highly vulnerable – remains diverse in terms of underlying needs and risk exposure. As a result, the intervention might be tailored

separately for more or less vulnerable youth, with separate treatment arms evaluating each formulation – so targeted scaling based on heterogeneous intervention impacts [List 2026]. This would of course layer on its own complications of fidelity and implementation for LA staff, and might be hard to reconcile with current norms of fixed teams of practitioners delivering either treatment or control to young people assigned to them. While an alternative might be to use incentives to ensure recruitment of the most vulnerable, it remains unclear who these incentives should be targeted to, given the flow of referrals is taken as given by Local Authorities. Such an approach might also raise ethical concerns over undue pressure being applied to vulnerable adolescents. A more indirect approach would be to provide LA-wide incentives to reach target quotas of the most vulnerable – similar to the design of clinical trials.

Second, our evidence suggests uncertainty over future funding can reduce buy-in and the perceived sustainability of interventions – so incorporating supply-side investments to help maintain voltage [List 2026]. The qualitative and cost data suggests sustaining voltage requires reductions in staff turnover – a friction that prevents gains from specialization being realized, and the cost data tentatively suggests can drive up per participant intervention costs at-scale. A natural option-C approach to quantitatively assess the extent of such challenges would be to vary the degree of uncertainty implementers face at scale-up, or to bundle interventions with incentives for the retention of key staff.

Finally, we note that a more general barrier to the adoption of option-C testing is the cost of such designs. While these might be lowered through coordinated efforts of funders and researchers to invest in shared research and development infrastructure [List 2024], in the short run we advocate for more combined use of quantitative and qualitative data collection exercises, to increase knowledge of the reasons for voltage drops. Many of the typical causes relate to system-wide factors – the emergence of complexities and uncertainties at-scale, the gradual loss of buy-in etc. Many of these are hard to measure quantitatively or are unrelated to the behavior of participants and delivery agents themselves, yet remain critical for understanding challenges to scaling, especially in organizations operating at near full capacity pre-intervention. It is perhaps through the use of such mixed methods approaches that more can be learned in the short term on the science of scaling, with the ultimate aim of informing policy makers of the effectiveness of interventions at-scale.[30]

[30] Blattman *et al.* [2017] and Abdul-Razzah *et al.* [2025] provide other examples of the combined use of such data, although their qualitative data focuses on establishing mechanisms of intervention impacts, rather than studying conditions for scaling.

A Appendix

A.1 Checkpoint Scale

The evaluation team, project team and the YPAG co-developed an instrument to measure adolescents' perceptions of their safety, named *Checkpoint: A Safety Scale for Adolescents* [Goodman *et al.* 2022]. The need to develop this arose from concerns in the project team and the YPAG about the acceptability of administering the Self-Report Delinquency Scale (SRDS) to the target cohort because the SRDS is not aligned with principles of the 'child-first' philosophy, principles of the YC intervention, and principles for engaging under-represented groups in an evaluation. Feedback from the YPAG and practitioners was that the SRDS could harm study participants due to the risk of stigmatizing them as criminals through questions about perpetration of criminal activities, and so likely to elicit inaccurate responses, and disrupt young person-practitioner relationships. Young people from marginalized groups were considered to be at most at risk of experiencing such harm, given their higher likelihood of such experiences and low levels of institutional trust.

The new scale was therefore developed with an objective to lower the potential stigmatization associated with asking young people about assumed anti-social and offending behaviors. Feedback suggested adolescents would be cautious about disclosing past and specific behaviors for fear of reprisal. This was addressed through the *Checkpoint* scale by asking them about their perception of risky behaviors rather than historic or specific behavior. To develop individual items, we reviewed questions in the SRDS and other self-reported measures of anti-social or offending behavior and attitudes to identify categories to measure: anti-social behavior, offending behavior, harm to others and harm from others (including exploitation). We co-developed six questions asking about the young person's likelihood to engage in these behaviors and/or be in harmful situations in the near future. While more acceptable to young people and to LAs, there are limitations in using *Checkpoint* as a measure of offending behavior. The instrument measures perceptions around safety, that is distinct from actual behavior.[31]

Validation We investigate the internal and external consistency of the self-reported and practitioner version of *Checkpoint*. The Cronbach alpha is .81 for the self-reported version and .89 for the practitioner version, both higher than the .7 threshold used to indicate satisfactory internal consistency. To assess external validity, Panel A of Figure A1 reports bivariate correlations between characteristics of the young person and their self-reported *Checkpoint* score. Panel B does the same for the practitioner-reported *Checkpoint* score. Our measure of self-reported risk is significantly lower for those aged 11-16, and significantly higher for those with disabilities and

[31] A safeguarding concern is raised when a young person reports it is likely/very likely that, in the next month, they will engage in behavior that could physically hurt others and/or that other people in their life will try and involve them in antisocial/violent/criminal behaviors, and/or that they will be a victim of a violent crime. Safeguarding concerns were raised for 20% of young people in the pilot.

from a white background. Practitioner assessments are significantly higher for male youth, those with special education needs, disabilities, and those whose most relevant primary need is socially unacceptable behavior. The last lines in each Panel show the correlations between the scores and the SDQ measures. Self-reported assessments are positively correlated with all measures and negatively correlated with the prosocial score. As expected, these correlations are stronger with the externalizing scores (conduct problems and hyperactivity) than with internalizing scores (emotional and peer problems). The pattern of bivariate correlations is similar for practitioners' risk assessment, with the exception of the internalizing scores, which are far less strongly positively associated with the risk assessment (and neither is statistically significant).

A.2 Qualitative Data Protocols

The research team jointly developed topic guides for staff and young persons based on our theory of change. The YPAG reviewed the initial draft of the young person topic guide and proposed new prompts and changes to wording of questions. The topic guide was revisited to help overcome challenges encountered during interviews. Interviews with young people took place on *Microsoft Teams* or in-person at a local authority hub, depending on the young person's choice. Interviews with staff took place on *Microsoft Teams*. A safeguarding contact for the young person at the LA was identified before interview. Participants were notified in the information sheet and again at the start of the interview that their responses were anonymous and confidential unless there was a risk of harm to self or others, in which case we would have to break confidentiality and contact their practitioner.

A.3 Data Validation in the Pilot

SDQ Measures To assess the quality of SDQ information collected at baseline, we compare the distributions of difficulty scores in our sample to known population thresholds and the Millennium Cohort Study (MCS). The MCS is a nationally representative survey of children born in 2001-02 across the UK, tracked since they were 9 months old. We focus on the age 14 MCS survey, given the age range of our study sample. Panel A of Figure A2 plots a histogram of the total difficulties score – the sum of all subscale items except the prosocial scale. This ranges from 0 to 40 and is counted as missing if one of the four component scores is missing. The distribution of the score is heavily shifted to the right in our pilot sample relative to the MCS, indicating a higher incidence of externalizing and internalizing problems than in a representative sample of 14 year olds. Panel B plots the histogram of prosocial scores. As expected, the distribution of prosocial scores is heavily skewed to the left in our pilot sample relative to the MCS.

While the SDQ is frequently used as a continuous score, it is also convenient to categorize these scores. The typical bandings used for SDQ scores are average, borderline and high/very high,

where these bandings are defined based on a population-based UK survey – the British Child and Adolescent Surveys of 1999+2004, where cut points are chosen such that 80% of children score average, 10% score borderline, and 10% score in the high/very high range. Panel C of Figure A2 compares the proportion of children in each category in our sample, the population-based survey, and the MCS. On the total difficulties score, the proportion of our sample in the borderline and high/very high range is over twice as high as expected based on the thresholds and four times as high as in the MCS age 14 sample. This is driven by the high frequency of externalizing problems: our pilot sample is 3.4 times more likely to be in the high/very high range of conduct problems and four times more likely to be in the high/very high range of hyperactivity than expected in the UK-wide population of 11-17 year olds.[32]

Finally we can also examine the relationships across subdomains. Table A22 reports correlations between the SDQ subdomains in our pilot sample and the MCS. The correlations are very similar (and also similar in the scale-up), further underpinning confidence in the quality of the self-reported SDQ measures.

A.4 Learnings from the Pilot

Data Collection Practices To improve the completion of session forms, a mobile-friendly online interface was developed to streamline practitioners' time costs of form completion and facilitate oversight by SPoCs. To address delays in the collection of endline surveys, the survey was revised to occur within regular sessions, supervised by practitioners and supplemented by audio assistance for participants facing reading challenges along with the option for surveys to be sent via a secure *WhatsApp* message. This contrasts with the approach used in the pilot which involved contacting participants outside regular sessions. To reduce attrition, we sought to understand risks of disengagement of control teams. Control practitioners' often perceived study participation to offer few immediate benefits to them and the young people they supported. Hence building up to and during the scale-up, considerable attention was given to ensuring practitioners in control teams felt integral to the evaluation, actively participated, and were recognized as stakeholders in the trial's success.

Assignment of Young People to Teams Fairness and practical concerns raised by LAs ensured we retained the same assignment procedure in the scale-up, based on team capacity when the young person is referred to and assessed by the LA. While biased assignment of adolescents to teams was not evident from baseline data or the qualitative interviews to allocators, we launched a large training campaign ahead of the scale-up, to remind all those involved about assignment, delivery and data collection protocols and the importance of following them strictly.

[32] The MCS sample tends to have higher proportions of children in the normal range than expected, which reflects selective attrition in MCS from the 9 month to the age 14 sample.

Design Issues YC was originally designed to involve 12 weeks of delivery with three weekly sessions of 45-60 minutes each. Without compromising the perceived adaptability of the intervention or reducing the overall intensity of treatment, the design was reformulated to require treated young people to meet their practitioner 36 times over 12 to 18 weeks. The timing of data collection in the scale-up was restructured to cover this extended timeline.

A.5 Power Calculations at Scale-up

To determine the minimum detectable effect size at scale-up, we elicited information from LAs about their recruitment expectations – the number of teams per LA, and the number of young people to be recruited per team. We assumed: (i) the control mean of being in the high/very high range of conduct problems would be similar at scale-up (30%) to the pilot (37%) given a similar set of eligible youth were expected to be recruited; (ii) attrition would fall to 30%, given modifications in data collection protocols; (iii) an intra-cluster correlation of .09. The calculations suggested we required 1761 young people across 132 teams to detect a minimum effect size of a 24% reduction in our primary outcome. The pilot showed the likelihood of high/very high conduct problems to fall by 33% of the baseline mean in controls. Moreover, prior CBT-based interventions on similar populations have achieved larger impacts on outcomes related to externalizing behaviors such as conduct problems, albeit with treatments delivered by specialist therapists under controlled conditions, to younger cohorts, and usually in a clinical environment [Riise *et al.* 2021]. While the sample size fell short of what was assumed to detect a minimum effect size of a 24% reduction in our primary outcome, two countervailing forces are relevant: attrition fell to 27% (rather than the assumed 30%) and the actual intra-cluster correlation for our primary outcome was .07 (rather than the assumed .09).

A.6 Dosage Effects at Scale-up

The failure to deliver YC with the intended intensity of 36 sessions partly reflects built-in flexibility that encouraged practitioners to adapt delivery to the needs of individual young persons, but also reflects system-wide constraints on practitioners. We examine whether voltage drops might have been prevented had the intervention been delivered with greater intensity by estimating dosage effects of YC in two ways: considering whether the young person received any YC sessions, and the total number of YC sessions received. We use an IV approach to instrument these outcomes using treatment assignment. Table A23 reports the coefficient on treatment assignment in the first stage, for both measures of dosage. As expected given the session forms evidence, the impact of treatment assignment on dosage is positive and statistically significant (p-value $< .001$) irrespective of the measure. Table A24 reports treatment effects in the second stage. Panel A shows the estimates from having at least one YC session, and Panel B shows the estimates from an additional session.

Nearly all results remain imprecise, and the point estimates of having at least one session in Panel A are often larger (in absolute value) than the baseline ITT estimates from Table 8, with the marginal effect of having an additional session being very small for most outcomes. Taken together the results reconfirm that any voltage drop is unlikely to be generated from the intensity of YC delivery falling off between the pilot and scale-up. These results align with existing meta-analysis showing that more intensive CBT-based treatments may be more effective, but the magnitudes of any dosage effects are small [Riise *et al.* 2021].

B Additional Qualitative Evidence

B.1 Pilot

B.1.1 Adolescents Assigned to Treatment

We summarize recurring themes from interviews during the pilot in relation to: (i) the view that YC helped build trusting adolescent-practitioner relationships; (ii) session content reflected the use of CBT-techniques; (iii) the theory of change was supported. All this underlines the promise of YC at the pilot stage.

Building Relationships Young people spoke positively about their practitioner, feeling practitioners understood or tried to get to know them, that in turn facilitated their session attendance. Adolescents indicated they could '*be open and speak my mind*' and '*feel free and safe to talk [to practitioner] about anything.*' Listening skills of practitioners paired with a helping attitude seemed to facilitate trust. As one adolescent said, '*When I first was coming, I wasn't really comfortable and stuff like that, and I would keep things to myself that I could have told [practitioner]. But when I first told [practitioner] something and I saw that [practitioner] was helping me, that [practitioner] was willing to help me with stuff, it made me feel more comfortable over time, to a point where I was just coming and telling [practitioner] all my situations and stuff.*'

Session Content Reflecting on behavior and consequences of actions was described as a repeated focus of sessions. Young people mentioned that techniques learned that helped them change behavior, such as '*[de-]catastrophising*' and '*consequential thinking.*' There was also an indication of changes in mindset as young people reflected on having '*a different view of seeing things in general*' or feeling more willing to take advice. Sessions helped young people to develop problem solutions, with one describing this as '*building myself and what issues I face with myself and how I should resolve that.*' Some mentioned working on goals outside of sessions, and all mentioned talking about jobs or '*what I want to do in life*' with their practitioner.

Theory of Change Having someone to talk to about feelings was perceived to be helpful to wellbeing: '*even if I'm having a hard day and I have mentoring that day, I'll still go because it's like I can vent, get stuff from my chest [...] before I feel like I couldn't really vent.*'; '*I'd say it feels like a safe space, like I'm wanting to come here and talk about something [...] even if [practitioner] can't help me, just to get it out.*' Young people reported increased school attendance, career ambitions, developing skills for employment, improvements to mental health, confidence and self-esteem, reductions in risky behaviors, and disengaging with peers involved in risky behaviors: '*[Working with practitioner] really opened my eyes about who you surround yourself with makes an image of who you are [...] I was chilling with people who were obviously missing school, misbehaving, had involvement with drugs, weapons, stuff like that, and the group I'm chilling with now are all people who are trying to be things.*'

B.1.2 Adolescents Assigned to Control

Building Relationships Those with more contact felt their practitioner tried to get to know them by asking about interests and meeting their family. However, not all young people felt like they could, or needed to, share information. For one young person the barrier was a lack of relationship leading them to feel uncomfortable, while another did not see a need: '*I don't need to tell [practitioner] anything because I'm not at risk.*'

Session Content Session activities in controls focused on education and employment, often in connection to goals. Practitioners supported young people by recommending training opportunities, communicating with firms, or working with schools/LAs to help them transition to a new school. As expected, there was some overlap in session content with that under YC. For example, as with YC, practitioners engaged young people in '*consequential thinking*', and how to react differently in situations: '*Before I do anything, I think about what the consequences are.*' Some young people reported a health/wellbeing focus to sessions: '*I don't really have problems with my mental state [...] but they still check up on me and ask me if everything's fine.*'; '*It's just asking questions and ticking what the good thing is, what's healthy and what's not healthy.*'

B.1.3 Practitioners' Views on Your Choice

Building Relationships Practitioners reported meeting with young people at least twice weekly, and described the intensity and flexibility of activities in YC as being important for building meaningful relationships: '*To work in quite a different way that also was within the remit of the work that I generally do was fantastic and it did transform my working relationship with that young person. It strengthened it...I got to see [them] at times in quite a different environment, relaxed, interacting differently. I think [they] saw a different side of me at times as well.*'

Practitioners indicated young people could '*see that you're wanting to change their behavior and support them to change their behavior*'. The consent-based nature of the intervention and voluntary participation helped create climates in which young persons did not feel pressured to talk about specific topics, fostering trust: '*I think that's one of the real strong points of the Your Choice programme. Those relationship building sessions aren't just an add-on, I think they're fundamental to making the Your Choice programme different to some of the interventions that they've had previously. [...] And these young people, often who are already very disengaged from the system, they don't trust adults, they certainly don't trust professionals. So actually to have an adult come alongside them and say, "You're important enough for me to actually just want to spend time with you and do something that is enjoyable, and get to know you."*'

Session Content Practitioners liked the focus on goals, thoughts and feelings rather than just presenting behaviors. They also described their use of CBT strategies, such as talking about feelings, challenging thoughts together, using exposure therapy and calming strategies.

Theory of Change Practitioners described a range of positive impacts of YC on young people, seeing improvements in their mental health ('*by accessing positive activities*'), family and relationships, confidence and self-esteem ('*through being able to identify additional needs around communication and putting things in place*'). They indicated increases in young people's safety by reduced exploitation concerns, and that some were more engaged in employment-related activities. This was described by a small number of practitioners as feeling the young person would not '*come back*', indicating reduced future risks of exposure to violence.

B.2 Scale-up

B.2.1 Young Persons' and Practitioners' Views on Your Choice

Building Relationships Young person-practitioner relationships continued to be an important for whether YC was seen as helpful. Young people continued to enjoy the flexibility of having sessions in different locations. All this led them to be able to talk about anything with their practitioners, and feel listened to: '*I will tell her anything and I know that I can trust her.*'; '*Having someone like [practitioner] that you can relate to, it just gives you that open space to kind of just talk.*'

Young people mentioned getting advice from their YC practitioners about problems at school, home, and with friends. They appreciated that practitioners had given them guidance and help in relation to education and employment, including setting up work experience opportunities, helping them to apply for jobs and apprenticeships, and enrolling them into college: '*[Practitioner] helped me with my CV, we done my CV, she applied for me and put my email in and stuff so I get the*

emails, and then she just took me to the interview, like, with college as well, she literally sorted it all out for me, I didn't even have to do anything.'

Theory of Change: Mechanisms Practitioners described the positive impact YC had on the young people they worked with, noticing changes in young people's behavior, reflectiveness, ways of handling situations, and communication with family: '*You could see the benefits because, like I said, there was a reduction in their behavior points.*' Mechanisms behind change from practitioners' perspectives included young people being able to set and track their goals: '*Having the goal there and seeing the steps that she needed to take, I think that helped her.*'

Young people also described emotional and behavioral changes as a result of YC, such as learning to deal with anger and stress, feeling more positive and less depressed. They discussed changes to habits, from having nothing to do, to engaging in activities which gave them a sense of purpose: '*She's just helped me to become more stronger like, endurance and stuff. Just to like deal with stuff better because back then I couldn't really deal with anything without getting angry or upset. It helped to recognise those emotions or feelings or whatever.*'; '*Back then I was just very negative and pretty much depressed because there's was nothing really good going on in my life. But now I can see the bright side of things and stuff and I can look forward to it.*'

They also described feeling calmer and learning methods to manage anger, stress, and anxiety. These included breathing, meditation, and counting methods: '*Sometimes, when I feel my chest getting tight, I be using those breathing exercises. And it just all releases it at once. It just gives you a massive relief. Do you get me? And you just feel better and just feel a bit more free.*'

On peer problems and prosocial behaviors, young people described being more sociable and less shy or nervous in social situations following YC: '*I just got a bit more confidence, because I was talking to [practitioner] and I was having conversations, so I thought maybe I can try and have conversations with other people.*' Practitioners also set up social activities for them, which helped them meet other young people.

Theory of Change: Outcomes Young people mentioned practitioners helping them with schoolwork, arranging extra tuition in particular subjects, and giving them confidence to put their hand up in class: '*Because, she was telling me, if I try it and I like it then she thinks I will be more engaged in my lessons and stuff. And I do agree with her because after I started putting my hand up in class more then I felt more engaged and I felt like I knew what I was doing in class as well.*'

The YC sessions also helped young people to make different choices and reduce their involvement in crime and violence by encouraging them to think before acting and consider the outcome: '*It turned me away from those kind of situations at least. Having this is like a deterrent.*' It also helped them recognise people who might be a bad influence: '*I think twice before meeting the wrong people, I think twice about where I'm going.*'

However, as with the pilot, the qualitative evidence does not paint a uniformly positive picture.

There were some young people that reported not noticing much change in their lives since they had started working with their YC practitioner. They also noted factors hindering their progress towards reaching the goals set in YC sessions, including contextual circumstances that were outside their control or their thoughts and feelings: '*What's been getting in the way or been challenging? Usually, like, my thoughts. Because I get a bit, like, sad and I'm just like, 'Oh, I'm not going to do it anymore.*"

Sustained Changes The scale-up qualitative data involved re-interviewing 12 of the young people six months after YC had ended, shedding light on more persistent changes in perceptions of the intervention. At this stage some young people described: (i) sustained improvements in emotion regulation, wellbeing, and developing a prosocial identity; (ii) they were still better able to manage their emotions, such as anger and anxiety; (iii) they mentioned continuing to follow advice from their YC practitioners: '*I don't have to keep asking people, 'What should I do if I'm angry?', I just remember what [practitioner] said.*'

Young people also described continuing to get into less trouble, due to learning from practitioners about how to handle situations differently: '*He would tell me how to approach different situations, and now I just use it because I don't get in trouble now as much as I used to in certain situations.*' They also mentioned experiencing improvements in their confidence and motivation and trying to maintain healthy sleeping and eating habits. They also described having more positive experiences at school compared to before YC, such as attending more regularly. Factors contributing to these more positive experiences included getting into fewer fights with peers, trying not to get angry with or be rude to teachers, paying attention, engaging more, or feeling more focused. '*I'm not as rude. How I used to, I'd talk with a lot, and I'd be really rude towards them. And I actually wouldn't be the nicest. I'd shout at them and all of that. But I don't really do that.*'

Young people also described ways in which relationships with parents or siblings had improved since their YC sessions, including having fewer arguments and communicating more. They also referenced changes in social relationships, including arguing less with peers, talking more about their feelings, and developing social skills such as speaking in large groups: '*If you know how each other feels, then you don't really argue about it because you understand how people are feeling.*'

However, not all adolescents had experienced lasting changes after their YC sessions had ended. Some felt they needed sustained help to manage ongoing anger and anxiety issues and to further improve their confidence and motivation: '*I need someone to talk to.*' Others reported still having problems with teachers or with their behavior, or mentioned that they were no longer attending school or college, or still experiencing difficulties within their social networks, including getting involved in fights and not agreeing with their practitioner's advice regarding peer relationships: '*I just have to resort to using violence if they use it on me first.*'

References

[1] ABDUL-RAZZAK.N, B.DOMASH, K.HALLBERG AND C.PINTO POEHLS (2025) Longer-Term Impacts of a Youth Behavioral Science Intervention: Experimental Evidence from Chicago, mimeo, University of Chicago.

[2] ADUKIA.A, B.FEIGENBERG AND F.MOMENI (2025) "From Retributive to Restorative: An Alternative Approach to Justice in Schools," *American Economic Review* 115: 2722-54.

[3] AIZER.A, J.J.DOYLE JR. (2015) "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges," *Quarterly Journal of Economics* 130: 759-804.

[4] AL-UBAYDLI.O, .S.LEE, J.A.LIST, C.L.MACKEVICIUS AND D.SUSKIND (2021) "How Can Experiments Play a Greater Role in Public Policy? Twelve Proposals from an Economic Model of Scaling," *Behavioural Public Policy* 5: 2-49.

[5] AL-UBAYDLI.O, J.A.LIST, D.LORE AND D.SUSKIND (2017) "Scaling for Economists: Lessons from the Non-Adherence Problem in the Medical Literature," *Journal of Economic Perspectives* 31: 125-44.

[6] BECK.J.S (2011) *Cognitive Behavior Therapy: Basics and Beyond* (2nd ed.), The Guilford Press.

[7] BENNETT.D AND T.GIBBONS (2000) "Efficacy of Child-behavioral Interventions for Antisocial Behavior: A Meta-analysis," *Child Fam. Behav. Ther.* 22: 1-15.

[8] BHATT.M.P, J.GURYAN, J.LUDWIG AND A.K.SHAH (2021) Scope Challenges to Social Impact, NBER WP28406.

[9] BHATT.M.P, S.B.HELLER, M.KAPUSTIN, M.BERTRAND AND C.BLATTMAN (2024) "Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago," *Quarterly Journal of Economics* 139:1-56.

[10] BLATTMAN.C, S.CHASKEL, J.JAMISON AND M.SHERIDAN (2023) "Cognitive Behavior Therapy Reduces Crime and Violence over 10 Years: Experimental Evidence," *American Economic Review: Insights* 5: 527-45.

[11] BLATTMAN.C, J.C.JAMISON AND M.SHERIDAN (2017) "Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia," *American Economic Review* 107: 1165-206.

[12] BRAITHWAITE.J, K.CHURRUCA, J.C.LONG, L.A.ELLIS AND J.HERKES (2018) "When Complexity Science Meets Implementation Science: A Theoretical and Empirical Analysis of Systems Change," *BMC Medicine* 16.

[13] BRAUN.V AND V.CLARKE (2006) "Using Thematic Analysis in Psychology," *Qualitative Research in Psychology* 3: 77-101.

[14] CARROLL.C.D, J.OVERLAND AND D.N.WEIL (2000) "Saving and Growth with Habit Formation," *American Economic Review* 90: 341-55.

[15] CATTAN.S ET AL. (2023) Your Choice: A Cluster Randomised Controlled Trial of a CBT-Informed Violence Reduction Programme Pilot Trial (Interim Findings), Youth Endowment Fund Report.

[16] CATTAN.S ET AL. (2025) Your Choice: A Cluster Randomised Controlled Trial of a CBT-Informed Violence Reduction Programme (Efficacy Report), Youth Endowment Fund Report.

[17] CHASSANG.S (2010) "Building Routines: Learning, Cooperation, and the Dynamics of Incomplete Relational Contracts," *American Economic Review* 100: 448-65.

[18] CRENNA-JENNINGS.W AND J.HUTCHINSON (2020) *Access to Child and Adolescent Mental Health Services in 2019*, Education Policy Institute.

[19] DAVIS.J.M.V, J.GURYAN, K.HALLBERG AND J.LUDWIG (2017) The Economics of Scale-up, NBER WP23925.

[20] DE CHAISEMARTIN.C AND J.RAMIREZ-CUELLAR (2024) "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?" *American Economic Journal: Applied Economics* 16: 193-212.

[21] DE HAAN.ANKE ET AL. (2024) "Efficacy and Moderators of Efficacy of Cognitive Behavioural Therapies with a Trauma Focus in Children and Adolescents: An Individual Participant Data Meta-analysis of Randomised Trials," *Lancet Child & Adolescent Health* 8: 28-39.

[22] DEL VECCHIO.T AND K.D.O'LEARY (2004) "Effectiveness of Anger Treatments for Specific Anger Problems: A Meta-analytic Review," *Clinical Psychology Review* 24: 15-34.

[23] DUMORNAY.N.M ET AL. (2022) "Improved Emotion Regulation Following a Trauma-informed CBT-based Intervention Associates with Reduced Risk for Recidivism in Justice-involved Emerging Adults," *Frontiers in Psychiatry* 13: 951429.

[24] DURLAK.J, L.RUBIN AND R.KAHNG (2001) "Cognitive Behavioral Therapy for Children and Adolescents with Externalizing Problems," *J. Cogn. Psychother.* 15: 183-94.

[25] EREN.O AND N.MOCAN (2021) "Juvenile Punishment, High School Graduation, and Adult Crime: Evidence from Idiosyncratic Judge Harshness," *Review of Economics and Statistics* 103: 34-47.

[26] FAGAN.A.A ET AL. (2019) "Scaling up Evidence-based Interventions in US Public Systems to Prevent Behavioral Health Problems: Challenges and Opportunities," *Prevention Science* 20: 1147-68.

[27] GAFFNEY.H, D.JOLLIFFEE AND H.WHITE (2021) *Trauma Informed Care: Toolkit Technical Report*, London: Youth Endowment Fund.

[28] GOODMAN.R (1997) "The Strengths and Difficulties Questionnaire: A Research Note," *Journal of Child Psychology and Psychiatry* 38: 581-6.

[29] GOODMAN.K, S.CATTAN, J.EDBROOKE-CHILDS, I.GREGORY AND E.STAPLEY (2022) *Checkpoint Scale*, supported by the Youth Endowment Fund and The London VRU Young People's Action Group.

[30] GOTTFREDSON.D.C ET AL. (2015) "Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science: Next Generation," *Prevention Science* 16: 893-926.

[31] HAINMUELLER.J (2012) "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies," *Political Analysis* 20: 25-46.

[32] HECKMAN.J.J, J.STIXRUD AND S.URZUA (2006) "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics* 24: 411-82.

[33] HEEKS.M, S.REED, M.TAFSIRI AND S.PRINCE (2018) *The Economic and Social Costs of Crime* (2nd ed.) Home Office Research Report.

[34] HELLER.S.B, A.K.SHAH, J.GURYAN, J.LUDWIG, S.MULLAINATHAN AND H.A.POLLACK (2017) "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago," *Quarterly Journal of Economics* 132: 1-54.

[35] HJALMARSSON.R (2008) "Criminal Justice Involvement and High School Completion," *Journal of Urban Economics* 63: 613-30.

[36] IRWIN-ROGERS.K, A.MUTHOO AND L.BILLINGHAM (2020) Youth Violence Commission Final Report, Youth Violence Commission.

[37] KILBOURNE.A.M, M.S.NEUMANN, H.A.PINCUS, M.S.BAUER AND R.STALL (2007) "Implementing Evidence-based Interventions in Health Care: Application of the Replicating Effective Programs Framework," *Implementation Science* 2:42.

[38] KLEMPERER.P (1987) "Markets with Consumer Switching Costs," *Quarterly Journal of Economics* 102: 375-94.

[39] KOEHLER.J.A, F.LÖSEL, T.D.AKOENSI AND D.K.HUMPHREYS (2013) "A Systematic Review and Meta-analysis on the Effects of Young Offender Treatment Programs in Europe," *Journal of Experimental Criminology* 9: 19-43.

[40] LANDENBERGER.N.A AND M.W.LIPSEY (2005) "The Positive Effects of Cognitive-behavioral Programs for Offenders: A Meta-analysis of Factors Associated with Effective Treatment," *Journal of Experimental Criminology* 1: 451-76.

[41] LEIJENHORST.L, B.MOOR, Z.MACKS, S.ROMBOUTS, P.WESTENBERG AND A.CRONE (2020) "Adolescent Risky Decision-making: Neurocognitive Development of Reward and Control Regions," *NeuroImage* 51: 345-55.

[42] LEVITT.S.D AND J.A.LIST (2007) "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?," *Journal of Economic Perspectives* 21: 153-74.

[43] LIPSEY.M.W (1999) "Can Rehabilitative Programs Reduce the Recidivism of Juvenile Offenders? An Inquiry into the Effectiveness of Practical Programs," *Virginia Journal of Social Policy & the Law* 6: 611-42.

[44] LIPSEY.M.W, N.A.LANDENBERGER AND S.J.WILSON (2007) "Effects of Cognitive-Behavioral Programs for Criminal Offenders," *Campbell Systematic Reviews* 6.

[45] LIST.J.A (2022) *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*, Penguin Randomhouse.

[46] LIST.J.A (2024) "Optimally Generate Policy-based Evidence Before Scaling," *Nature* 626(7999): 491-9.

[47] LIST.J.A (2026) "The Economics of Scaling Early Childhood Programs: Lessons from the Chicago School," *Journal of Political Economy* 134: 1-48.

[48] LIST.J.A, A.M.SHAIKH AND A.VAYALINKAL (2023) "Multiple Testing with Covariate Adjustment in Experimental Economics," *Journal of Applied Econometrics* 38: 920-39.

[49] NG FAT.L, S.SCHOLES, S.BONIFACE, J.MINDELL AND S.STEWART-BROWN (2017) "Evaluating and Establishing the National Norms for Mental Well-being Using the Short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS): Findings from the Health Survey for England," *Quality of Life Research* 26: 1129-44.

[50] OWENS.E.G (2017) "Testing the School-to-prison Pipeline," *Journal of Policy Analysis and Management* 36: 11-37.

[51] QUITANGON.G (2019) "Vicarious Trauma in Clinicians: Fostering Resilience and Preventing Burnout," *Psychiatric Times* 36: 18-9.

[52] RIISE.E.N, G.J.H.WERGELAND, U.NJARDVIK AND L-G.OST (2021) "Cognitive Behavior Therapy for Externalizing Disorders in Children and Adolescents in Routine Clinical Care: A Systematic Review and Meta-analysis," *Clinical Psychology Review* 83: 101954.

[53] RITCHIE.J AND L.SPENCER (1994) "Qualitative Data Analysis for Applied Policy Research, in Analyzing Qualitative Data," in *Analyzing Qualitative Data*, A.Bryman and B.Burgess (eds.) Routledge: London.

[54] ROMANO.J.P AND M.WOLF (2005) "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica* 73: 1237-82.

[55] ROMANO.J.P, A.M.SHAIKH AND M.WOLF (2008) "Control of the False Discovery Rate Under Dependence Using the Bootstrap and Subsampling," *TEST* 17: 417-42.

[56] ROMANO.J.P AND M.WOLF (2016) "Efficient Computation of Adjusted p-values for Resampling-based Stepdown Testing," *Statistics and Probability Letters* 113: 38-40.

[57] ROSS.L AND R.E.NISBETT (2011) *The Person and the Situation: Perspectives of Social Psychology*, Pinter & Martin.

[58] SAINI.M (2009) "A Meta-analysis of the Psychological Treatment of Anger: Developing Guidelines for Evidence-based Practice," *Journal of the American Academy of Psychiatry and the Law* 37: 473-88.

[59] SCHONBERG.T, C.FOX, J.MUMFORD, E.CONGDON, C.TRAPEL AND R.POLDRACK (2012) "Decreasing Ventromedial Prefrontal Cortex Activity During Sequential Risk-taking: An FMRI Investigation of the Balloon Analog Risk Task," *Front. Neurosci.* 6: 1-11.

[60] SHEM-TOV.Y, S.RAPHAEL AND A.SKOG (2024) "Can Restorative Justice Conferencing Reduce Recidivism? Evidence From the Make-it-Right Program," *Econometrica* 92: 61-78.

[61] SPAAN.P, F.VAN DEN BOOGERT, N.H.GROOTENDORST-VAN MIL, W.J.G.HOOGENDIJK AND S.J.ROZA (2023) "Screening for Disruptive Behavior in Adolescents at Risk Using the Strengths and Difficulties Questionnaire," *Journal of Research on Adolescence* 33: 1085-97.

[62] STEINBERG.L (2008) "A Social Neuroscience Perspective on Adolescent Risk-Taking," *Dev Rev.* 28: 78-106.

[63] STEVENSON.M (2023) "Cause, Effect, and the Structure of the Social World," *Boston University Law Review* 103: 2001-47.

[64] WEISS.M.J, H.S.BLOOM AND T.BROCK (2014) "A Conceptual Framework for Studying the Sources of Variation in Program Effects," *Journal of Policy Analysis and Management* 33: 778-808.

[65] WENZEL.A (2021) "The Therapeutic Relationship," in A.Wenzel (ed.), *Handbook of Cognitive Behavioral Therapy: Overview and Approaches*, American Psychological Association.

[66] WICHMANN.M.L, S.PAWILS, J.RICHTERS AND F.METZNER (2023) "School-Based Interventions for Child and Adolescent Victims of Interpersonal Violence," *Trauma Violence Abuse* 24:1743-62.

[67] WILSON.D.B, L.A.BOUFFARD AND D.L.MACKENZIE (2015) "A Quantitative Review of Structured, Group-Oriented, Cognitive-Behavioral Programs forOffenders," *Criminal Justice and Behavior* 32: 172-204.

[68] YARDLEY.L (2000) "Dilemmas in Qualitative Health Research," *Psychology & Health* 15: 215-28.

[69] YJB (2024) *Youth Justice Statistics 2023-24*, Youth Justice Board.

Table 1: Background Characteristics at Baseline, Pilot

Mean, difference and p-values in brackets

	(1) Control	(2) Treatment	(3) Difference	(4) p-value
Number of observations	79	114		
A: Demographics				
Female	.302	.303	.002	[.983]
Aged 16/17 (vs. aged 11-15)	.544	.412	-.132	[.213]
White	.253	.237	-.016	[.872]
Black	.266	.246	-.020	[.803]
Asian	.139	.044	-.094	[.030]
Mixed race/ethnicity	.114	.167	.053	[.239]
Special education needs	.226	.291	.065	[.273]
Disability	.444	.489	.044	[.696]
In social care	.174	.151	-.023	[.760]
B: Primary Need Assessment				
Family dysfunction	.202	.193	-.010	[.888]
Socially unacceptable behavior	.139	.263	.124	[.083]
Family in acute stress	.076	.070	-.006	[.090]
Other	.291	.123	-.168	[.025]
Not stated	.291	.351	.060	[.490]
C: Activities				
School/college	.609	.745	.136	[.058]
Work	.072	.010	-.062	[.159]
Training	.072	.031	-.042	[.371]
Other	.246	.201	-.032	[.653]

Notes: The sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. We report means of each characteristic in control and treated groups (Columns 1 and 2), their raw difference (Column 3), and the p-value on a t-test of equality of means (Column 4). This is based on an OLS regression where standard errors are clustered at the Local Authority level. Panels A and B are derived from Local Authority Workbooks. Panel C is derived from the baseline survey to young persons.

Table 2: SDQ-based Outcomes at Baseline, Pilot

Mean, standard deviation, difference and p-values in brackets

	(1) Control	(2) Treatment	(3) Difference	(4) p-value
A. Primary Outcome				
Conduct problems score (0-10)	3.64	3.81	.169	[.637]
	(2.14)	(2.26)		
Conduct problems: high/very high range	.373	.400	.027	[.702]
B. Secondary Outcomes				
Hyperactivity score (0-10)	5.39	5.82	.436	[.428]
	(2.79)	(2.56)		
Hyperactivity score: high/very high range	.369	.421	.052	[.616]
Emotional problems score (0-10)	3.97	3.18	-.791	[.163]
	(2.51)	(2.51)		
Emotional problems: high/very high range	.288	.168	-.119	[.166]
Peer problems score (0-10)	2.97	2.61	-.358	[.306]
	(1.95)	(1.96)		
Peer problems: high/very high range	.382	.312	-.071	[.348]
Prosocial behaviors score (0-10)	6.99	6.74	-.242	[.428]
	(2.19)	(2.06)		
Prosocial behaviors: low/very low range	.212	.280	.067	[.275]
Social connectedness scale (1-5)	3.54	3.69	.148	[.287]
	(.881)	(.717)		

Notes: The sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. We report means (standard deviations) in control and treated groups (Columns 1 and 2), their raw difference (Column 3), and the p-value on a t-test of equality of means (Column 4). This is based on an OLS regression where standard errors are clustered at the Local Authority level. Panels A and B are derived from the baseline survey to young persons. The outcomes are subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). The Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere.

Table 3: *Checkpoint* Risk Assessment at Baseline, Pilot

Mean, difference and p-values in brackets

	(1) Control	(2) Treatment	(3) Difference	(4) p-value
A: Self-Assessed Risks				
Checkpoint safety score (1-5)	2.09	2.00	-.086	[.576]
	(.888)	(.762)		
In the next month, likely/very likely to:				
to cause damage in public place	.118	.116	-.002	[.968]
to get in trouble with police	.118	.126	.009	[.870]
to physically hurt someone	.118	.095	-.023	[.702]
others will involve in any of above	.134	.168	.034	[.509]
to witness a violent crime	.239	.179	-.060	[.350]
to be victim of violent crime	.076	.043	-.033	[.435]
B: Practitioner Assessed Risks				
Checkpoint safety score (1-5)	2.86	3.10	.249	[.161]
	(.821)	(.761)		
In the next month, likely/very likely to:				
to cause damage in public place	.321	.261	-.061	[.486]
to get in trouble with police	.360	.451	.091	[.299]
to physically hurt someone	.200	.264	.064	[.352]
others will involve in any of above	.375	.696	.321	[.000]
to witness a violent crime	.393	.424	.031	[.830]
to be victim of violent crime	.240	.244	.004	[.950]

Notes: The sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. We report means (standard deviations) in control and treated groups (Columns 1 and 2), their raw difference (Column 3), and the p-value on a t-test of equality of means (Column 4). This is based on an OLS regression where standard errors are clustered at the Local Authority level. Panel A is derived from the Checkpoint risk assessment tool administered as part of the baseline survey to young persons. Panel B is derived from the Checkpoint risk assessment tool administered as part of baseline survey to practitioners.

Table 4: Treatment Effects on Primary and Secondary Outcomes, Pilot

OLS estimates, standard errors clustered at the local authority level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness	Emotional Self-regulation	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Your Choice	-.105	-.084	.051	-.038	.001	.076	.014	.076	-.052	.149
	(.094)	(.102)	(.102)	(.148)	(.071)	(.147)	(.205)	(.204)	(.179)	(.188)
p-val	*[.280]*	*[.422]*	*[.624]*	*[.798]*	*[.993]*	*[.613]*	*[.946]*	*[.713]*	*[.777]*	*[.441]*
Baseline outcome	.334***	.504***	.571***	.560***	.494***	.400***			.403***	.417***
	(.107)	(.092)	(.094)	(.078)	(.078)	(.104)			(.107)	(.094)
p-val	*[.007]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.001]*			*[.002]*	*[.000]*
Control Mean Baseline	.333	.423	.308	.308	.154	3.81	-	-	2.28	2.89
Characteristics	No	No	No	No	No	No	No	No	No	No
Observations	81	80	79	78	78	78	96	85	78	80
Clusters	29	29	28	28	29	29	33	30	29	30

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. OLS estimates are presented where standard errors are clustered at the Local Authority level, and p-values are reported in brackets. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 9 and 10 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table 5: Background Characteristics at Baseline, Scale-up

Mean, difference and p-values in brackets

	(1) Control	(2) Treatment	(3) Difference (within LA)	(4) p-value
Number of observations	611	854		
A: Demographics				
Female	.373	.288	-.028	[.584]
Aged 16/17 (vs. aged 11-15)	.568	.444	-.073	[.152]
White	.139	.201	-.017	[.542]
Black	.191	.204	.014	[.631]
Asian	.157	.080	-.034	[.165]
Mixed race/ethnicity	.257	.204	-.030	[.330]
Special education needs	.123	.241	.075	[.014]
Disability	.208	.306	.069	[.135]
In social care	.129	.162	-.005	[.920]
B: Primary Need Assessment				
Family dysfunction	.059	.091	.039	[.052]
Socially unacceptable behavior	.380	.384	-.034	[.484]
Family in acute stress	.052	.066	-.013	[.485]
Other	.332	.215	-.044	[.448]
Not stated	.177	.244	.051	[.144]
C: Activities				
School/college	.827	.678	-.090	[.129]
Work	.064	.037	-.019	[.192]
Training	.025	.049	.018	[.142]
Other	.127	.243	.081	[.155]

Notes: The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. We report means in control and treated groups (Columns 1 and 2), their within local authority difference (Column 3), and the p-value on the corresponding t-test of equality of means (Column 4). This is based on an OLS regression where standard errors are clustered at the Local Authority level. Panels A and B are derived from Local Authority Workbooks. Panel C is derived from the baseline survey to young persons.

Table 6: Primary and Secondary Outcomes at Baseline, Scale-up

Mean, standard deviation, difference and p-values in brackets

	(1) Control	(2) Treatment	(3) Difference (within LA)	(4) p-value
A. Primary Outcome				
Conduct problems score (0-10)	3.04	3.69	.370	[.075]
	(2.11)	(2.250		
Conduct problems: high/very high range	.247	.346	.038	[.395]
B. Secondary Outcomes				
Hyperactivity score (0-10)	5.01	5.59	.280	[.276]
	(2.66)	(2.62)		
Hyperactivity score: high/very high range	.308	.381	.027	[.578]
Emotional problems score (0-10)	3.12	3.08	.033	[.894]
	(2.35)	(2.32)		
Emotional problems: high/very high range	.192	.153	-.014	[.711]
Peer problems score (0-10)	2.72	3.01	.300	[.089]
	(1.78)	(1.85)		
Peer problems: high/very high range	.300	.348	.036	[.386]
Prosocial behaviors score (0-10)	7.27	6.85	-.307	[.042]
	(1.88)	(2.04)		
Prosocial behaviors: low/very low range	.181	.250	.048	[.122]
Social connectedness scale (1-5)	3.73	3.62	-.067	[.389]
	(.738)	(.749)		
Prosocial identity scale (1-5)	2.84	2.67	-.100	[.042]
	(.670)	(.684)		

Notes: The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. We report means (standard deviations) in control and treated groups (Columns 1 and 2), their within local authority difference (Column 3), and the p-value on the corresponding t-test of equality of means (Column 4). This is based on an OLS regression where standard errors are clustered at the Local Authority level. Panels A and B are derived from the baseline survey to young persons. The outcomes are subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). The Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. The prosocial identity is measured by the Prosocial Identity Scale.

Table 7: *Checkpoint* Risk Assessment at Baseline, Scale-up

Mean, difference and p-values in brackets

	(1) Control	(2) Treatment	(3) Difference (within LA)	(4) p-value
A: Self-Assessed Risks				
Checkpoint safety score (1-5)	1.67	1.87	.117	[.083]
	(.752)	(.870)		
In the next month, likely/very likely to:				
to cause damage in public place	.044	.099	.042	[.112]
to get in trouble with police	.035	.094	.048	[.045]
to physically hurt someone	.056	.090	.035	[.152]
others will involve in any of above	.098	.140	.039	[.111]
to witness a violent crime	.176	.188	-.011	[.684]
to be victim of violent crime	.073	.065	-.017	[.458]
B: Practitioner Assessed Risks				
Checkpoint safety score (1-5)	2.36	2.76	.353	[.003]
	(.841)	(.847)		
In the next month, likely/very likely to:				
to cause damage in public place	.116	.213	.092	[.012]
to get in trouble with police	.143	.318	.125	[.044]
to physically hurt someone	.096	.222	.105	[.037]
others will involve in any of above	.283	.453	.123	[.063]
to witness a violent crime	.186	.226	.058	[.253]
to be victim of violent crime	.090	.161	.086	[.017]

Notes: The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. We report means (standard) deviation in control and treated groups (Columns 1 and 2), their within local authority difference (Column 3), and the p-value on the corresponding t-test of equality of means (Column 4). This is based on an OLS regression where standard errors are clustered at the Local Authority level. Panel A is derived from the Checkpoint risk assessment tool administered as part of the baseline survey to young persons. Panel B is derived from the Checkpoint risk assessment tool administered as part of baseline survey to practitioners.

Table 8: Treatment Effects on Primary and Secondary Outcomes, Scale-up

OLS estimates, standard errors clustered at the LA level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness scale	Emotional Self-regulation	Prosocial identity scale	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Your Choice	.028	-.010	-.007	-.032	.025	.046	.056	-.017	-.025	-.062	.199
	(.029)	(.030)	(.021)	(.034)	(.038)	(.050)	(.066)	(.050)	(.074)	(.062)	(.145)
p-val (unadjusted)	*[.328]*	*[.734]*	*[.739]*	*[.352]*	*[.508]*	*[.369]*	*[.404]*	*[.741]*	*[.740]*	*[.321]*	*[.180]*
adj p-val (RW)		*[.862]*	*[.862]*	*[.599]*	*[.758]*	*[.563]*	*[.563]*	*[.860]*	*[.860]*	*[.136]*	*[.044]*
adj p-val (LSV)		*[.929]*	*[.737]*	*[.771]*	*[.853]*	*[.827]*	*[.777]*	*[.919]*	*[.740]*	*[.320]*	*[.434]*
Baseline outcome	.264***	.453***	.370***	.309***	.356***	.460***		.190***		.319***	.590***
	(.041)	(.028)	(.034)	(.036)	(.041)	(.035)		(.058)		(.045)	(.070)
p-val	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*		*[.003]*		*[.000]*	*[.000]*
H0: Effect Scale-up = Pilot											
p-val	[.000]	[.019]	[.011]	[.855]	[.519]	[.549]	[.531]		[.180]	[.862]	[.733]
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	962	962	957	963	967	994	1,066	983	1,024	961	1,091
Clusters	30	30	30	30	30	30	30	30	30	30	30

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. OLS estimates are presented where standard errors are clustered at the Local Authority level. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 prosocial identity is measured by the Prosocial Identity Scale. In Column 9 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 10 and 11 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available and Local Authority fixed effects. We report conventional p-values, p-values that adjust for multiple hypothesis testing using Romano Wolf algorithm (RW), and p-values that adjust for multiple hypothesis testing using List Shaikh Vayalinka (2021) approach (LSV). These account for MHT across Columns 2 to 5, 6 to 9, and 10 and 11. At the foot of each Column we report the p-value of a t-test of the null hypothesis that the treatment effect in the scale-up is equal to the value of the treatment effect in the pilot. We also report the mean outcome among controls tracked to endline.

Table 9: Sample Unrepresentativeness

OLS estimates, standard errors clustered at the LA level, p-values in brackets
Outcome: SDQ-scale conduct problems lie in the high/very high range

	Pilot	Scale-up	Services and young person characteristics	LASSO selected
	(1)	(2)	(3)	(4)
Your Choice	-.105	.028	-.160	-.106
	(.094)	(.029)	(.119)	(.158)
p-val	*[.280]*	*[.328]*	*[.190]*	*[.509]*
Baseline outcome	.334***	.264***	.179	.189
	(.107)	(.041)	(.135)	(.127)
p-val	*[.007]*	*[.000]*	*[.196]*	*[.149]*
Control Mean Baseline (weighted)	.333	.246	.433	.378
Strata FE (LA)	No	Yes	Yes	Yes
Observations	81	962	810	843

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. In Column 3, the scale-up sample is reweighted using entropy weighting such that the first moments of the demographic characteristics, primary need, service indicators, and baseline outcomes match those of the pilot sample. In Column 4, the reweighting is done on a subset of these variables selected by Lasso as the best predictors of being in the pilot sample, which includes indicators for white, black, special education needs, disability, in social care, the primary need assessment being family dysfunction, the primary need assessment being family stress, the primary need assessessment as being not stated, baseline conduct problems in the high/very high range, baseline emotional problems in the high/very high range, the practitioner-assessed *Checkpoint* safety score, and the service type -- Children Looked After, Early Help, and Youth Justice. OLS estimates are presented where standard errors are clustered at the Local Authority level. The outcome is based on the conduct problems subscale from the Strengths and Difficulties Questionnaire, and is an indicator of whether the individual scores in the high/very high range. At the foot of each Column we report the (weighted) mean outcome among controls tracked to endline.

Table 10: Heterogeneity of Impacts by Baseline Outcome, Scale-up

OLS estimates, standard errors clustered at the local authority level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness scale	Prosocial identity scale	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Your Choice: Problematic range at BL	-.062	-.135	.014	-.074	.093	.229	-.054	-.286	.407
	(.043)	(.044)	(.068)	(.066)	(.091)	(.109)	(.083)	(.140)	(.239)
p-val	*[.156]*	*[.005]*	*[.841]*	*[.274]*	*[.314]*	*[.044]*	*[.524]*	*[.051]*	*[.099]*
Your Choice: Normal range at BL	.073	.076*	-.012	-.012	.008	-.021	-.004	.015	.242**
	(.047)	(.044)	(.020)	(.035)	(.045)	(.044)	(.061)	(.069)	(.116)
p-val	*[.127]*	*[.093]*	*[.577]*	*[.739]*	*[.853]*	*[.634]*	*[.951]*	*[.826]*	*[.046]*
Difference	-.136*	-.211***	.025	-.062	.085	.251**	-.050	-.301*	.165
	(.069)	(.069)	(.070)	(.070)	(.106)	(.100)	(.101)	(.164)	(.205)
p-val	*[.059]*	*[.005]*	*[.719]*	*[.386]*	*[.430]*	*[.018]*	*[.624]*	*[.077]*	*[.428]*
Definition of problematic range at BL	High/very high conduct problems	High/very high hyperactivity	High/very high emotional problems	High/very high peer problems	Low/very low prosocial behaviours	Bottom 30% on social connectedness	Bottom 30% on prosocial identify scale	Bottom 30% on Checkpoint Self-assessed	Bottom 30% on Checkpoint Practitioner-Assessed
Mean Outcome Normal range at BL	.086	.087	.109	.220	.168	3.95	3.73	1.42	1.59
Mean Outcome Problematic range at BL	.455	.684	.472	.587	.475	3.19	3.45	2.10	2.53
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	962	962	957	963	967	994	983	961	1,091

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. OLS estimates are presented where standard errors are clustered at the Local Authority level. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 prosocial identity is measured by the Prosocial Identity Scale. Columns 8 and 9 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control for an indicator of the baseline outcome being in a `problematic range', an interaction between this indicator and treatment, and Local Authority fixed effects. We report conventional p-values. At the foot of each Column we report the exact definition of the problematic range for each baseline outcome. Note that for all SDQ outcomes, this corresponds exactly to the baseline outcome. We also report mean outcome among controls tracked to endline in the two groups.

Table 11: Costs

Cost category and item	Row	Pilot (1) Pilot	Pilot (2) Per participant cost	Scale-up (3) Scale-up	Scale-up (4) Per participant cost	(5) Ratio of per participant costs
Mean number of young people treated by a typical Your Choice team in a given month			2.7		3.8	
A. Fixed Costs						
Operational lead						
Data and governance activities	1	£1,960	£726	£1,280	£337	.464
Attendance of operational leads	2	£245	£91	£545	£143	1.58
Familiarisation with Your Choice	3	£3,367	£1,247	£3,364	£885	.710
Overview of cases and supervision of practitioners	4	£1,002	£371	£170	£45	.121
Clinical lead						
Attendance of clinical leads	5	£293	£109	£787	£207	1.91
Delivery of training to practitioners	6	£1,368	£506	£4,249	£1,118	2.21
Participation in training (delivered by LIIA)	7	£1,368	£506	£1,151	£303	.598
Additional data and governance activities	8	£630	£233	£2,648	£697	2.99
Practitioners						
Attendance of coach network groups	9	£1,044	£387	£690	£182	.470
Participation in training (delivered by trainer)	10	£1,044	£387	£3,980	£1,047	2.71
Overview of cases and supervision of practitioners	11	£843	£312	£1,654	£435	1.39
Materials and equipment						
Tablets or any other devices for practitioners	12			£2,973	£782	-
Total Fixed Costs	13	**£13,164**	**£4,875**	**£23,491**	**£6,182**	
B. Recurring Costs						
Clinical lead						
Provision of clinical supervision to practitioners	14	£593	£220	£10,507	£2,765	12.6
Practitioners						
Delivery of Your Choice – practitioner preparation ahead of sessions and case recordings	15	£2,732	£1,012	£499	£131	.130
Delivery of Your Choice – practitioners' attendance of clinical supervision meetings	16	£342	£126	£5,597	£1,473	11.6
Delivery of Your Choice– conducting sessions with young people (and families/teachers if appropriate)	17	£6,375	£2,361	£16,766	£4,412	1.87
Materials and equipment						
Travel	18	£481	£178	£978	£257	1.44
Behavioral activation funds (engagement activities)	19	£914	£338	£2,140	£563	1.66
Step-down costs (to enable programme impact to be sustained for young people)	20	£135	£50	£5,250	£1,382	27.6
Total Recurring Costs	21	**£11,573**	**£4,286**	**£41,737**	**£10,983**	
Total Cost	22	**£24,736**	**£9,162**	**£65,228**	**£17,165**	

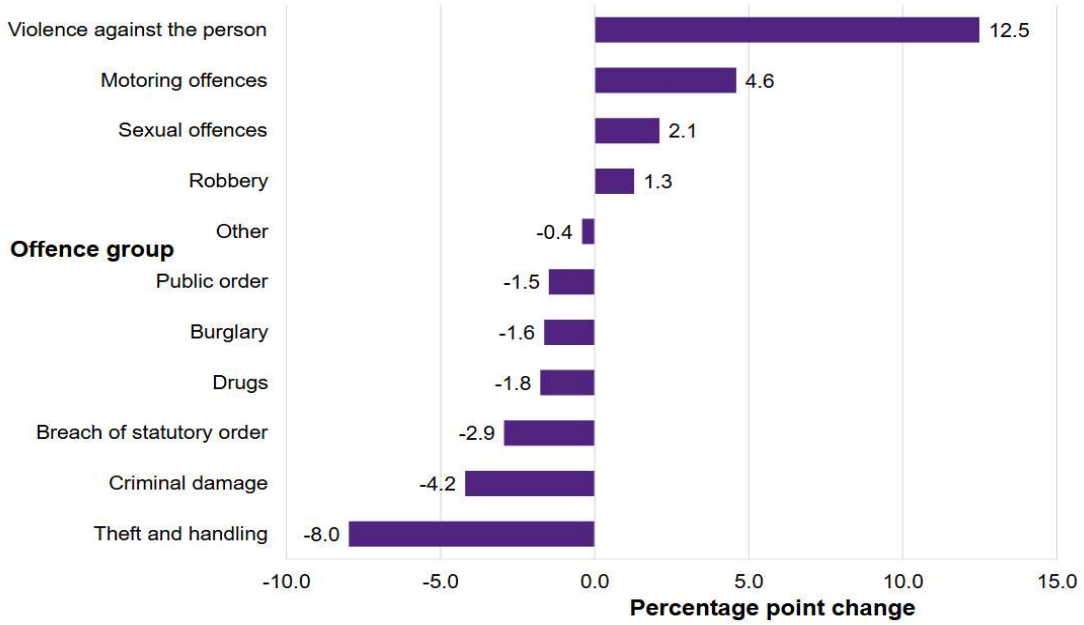
Notes: In the pilot phase, one Local Authoritiy provided suficiently detailed cost data. In the scale-up phase, seven Local Authorities did so - and we take their average reports of each line cost reported.

Table 12: Challenges to Scaling

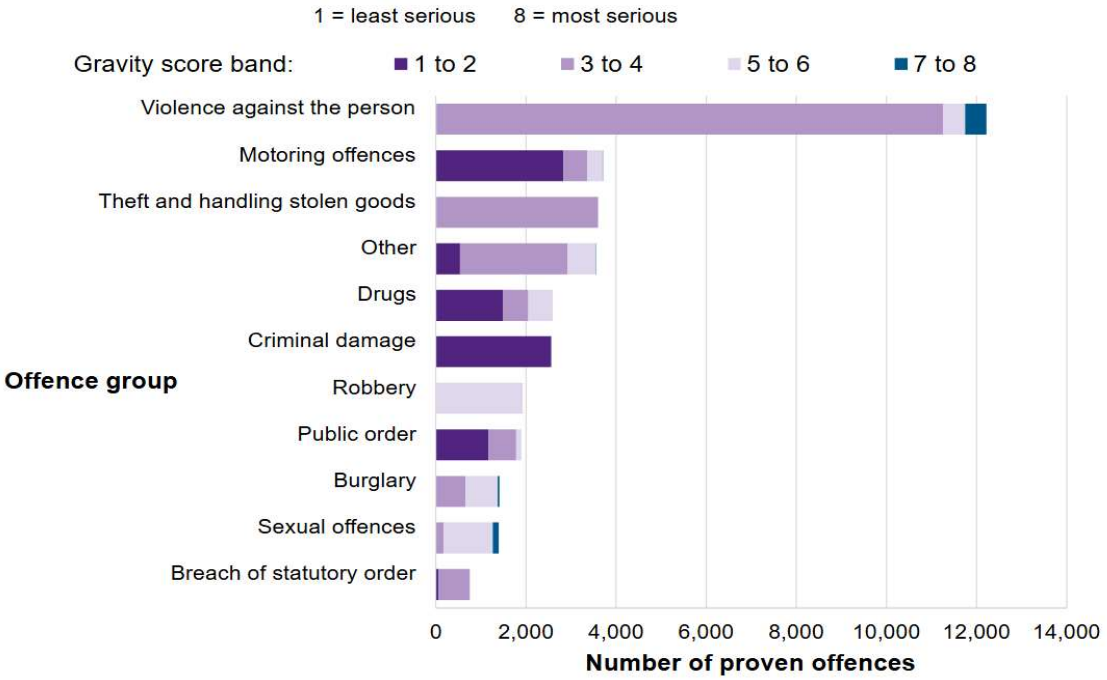
Mechanism generating voltage drop	Driving voltage drop?	Evidence	Quantitative-qualitative evidence aligned?
1. False positive	**Unlikely**	**Quantitative:** Pilot samples are balanced for SDQ-outcomes, other secondary outcomes and risk assessments (Tables 2, 3). Pilot results show the direction and size of point estimate treatment effects on the primary outcome shifting in line with the theory of change, through the intended mechanisms as described in the theory of change (Table 4). Sessions data for the pilot show the greater number of sessions for treated youth than controls, and the greater use of CBT-techniques within treated sessions (Section 3.3). **Qualitative:** Interviews with young persons and practitioners suggest elements of the theory of change were activated and the intervention had the intended shift on outcomes. Interviews with senior LA staff responsible for the assignment of young persons to teams also suggests the research design was adhered to (Section 3.5).	Yes
2. Sample unrepresentativeness	**Very likely**	**Quantitative:** The pilot sample presents worse characteristics at baseline relative to those recruited during scale-up: they are more likely to have special education needs or a disability, more likely to be in the extreme range of conduct problems, emotional problems, and self-assess to have greater risk of exposure to violence (Table A18 to A20). Reweighting the scale-up sample to match baseline characteristics of those in the pilot, we can nearly replicate the pilot impacts for the primary outcome, and the reweighted treatment effects on some secondary outcomes also move towards those found in the pilot (Table 9, Table A21). Estimating heterogeneous impacts by baseline SDQ and risk assessment scores, we confirm the impacts of YC are larger for those in the problematic range for conduct problems, hyperactivity, and self-assessed risks (Table 10).	N/A (we only use quantitative evidence)
	Unlikely	**Quantitative: T**he motivation of practitioners is unlikely to have decreased in scale up relative to pilot - if anything it might move in the opposite direction and increase voltage given that in the scale-up, practitioners assess treated young persons to be of higher risk compared to controls, than this difference in practitioner reports in the pilot (Tables 3 and 7).	
3. Situational unrepresentativeness	**Unlikely**	**Quantitative:** YC continued to be delivered as designed. The number of sessions increased during scale-up relative to the pilot, and session content continued to show the delivery of CBT techniques and building trusting relationships (Section 4.3). Supported by evidence from fidelity indices collected from young persons and practitioners (Table A9). Suggests neither the quality of implementation unlikely nor compliance with the intervention changed at-scale. **Qualitative:** Interviews with young persons and practitioners during scale-up reaffirms what was previously documented during the pilot: (i) young people viewed relationship building as key to the success of YC, enabling them to talk openly to practitioners; (ii) young people and practitioners reported the use of CBT-tools in YC sessions; (iii) the underlying theory of change related to mechanisms and intended outcomes continued to be supported from both perspectives, with evidence of changes being sustained over time for young people (Section 4.5).	Mostly
	Likely	**Qualitative:** Flexibility of the intervention still seen as a key asset and one of the reasons practitioners liked the intervention. However flexibility in delivery means there is no single recipe to be passed on, and more experienced practitioners may be better equipped to ensure successful delivery to adolescents. This is problematic for an intervention that relies on the accumulated and combined experiences of practitioners, trainers and clinical supervisors, but operates within real world constraints, with high staff turnover.	
4. Supply-side constraints	**Likely**	**Qualitative:** Resistance to change and buy-in were perhaps higher in the pilot than the scale-up, although uncertainty over future funding reduced buy-in over the scale-up. The intervention gained practitioner buy-in because it was perceived to add to their toolkit, providing opportunities for professional development, and bring structure to their work and the young people they worked with. However, a key uncertainty that was only realized at-scale is staff turnover. Staff turnover means that young people's involvement in YC could end if their practitioner left, or services could lose entire practitioner teams. Turnover of clinical leads led to particularly acute strains on voltage because of the consequent reductions in clinical supervision, and a reduced ability to train incoming practitioners. A third complexity that emerged at-scale relatives to workloads. Operating at scale increased the demands on capacity as practitioners were required to recruit and deliver the intervention to more young people (Section 5.3).	N/A (we only use qualitative evidence)
5. Cost-inflation	**Likely (but tentative)**	**Quantitative:** This is exploratory analysis given a limited number of LAs that could provide such data. It suggests the emergence of diseconomies of scale: the average cost per participant almost doubled from the pilot to scale-up because: (i) higher costs related to the clinical supervision of practitioners -- that partly reflects staff turnover among senior leads preventing gains from specialization to be fully realized; (ii) costs incurred as local authorities provided resources to enable young people to continue with elements of YC to sustain its impact, despite the lack of future funding guarantees.	N/A (we only use qualitative evidence)
6. General equilibrium effects	**Unlikely**	**Conjectural:** We do not have direct evidence on this, but increased workloads and costs related to YC might have had negative spillovers to youth in other parts of adolescent services. This concern is ameliorated by fact that practitioners work within the same team -- they do not typically divide their time between teams. Similarly, all youth assigned to a team have the same treatment status -- so there is no margin for treated practitioners to reallocate effort between control and treated adolescents.	

Figure 1: Youth Justice Statistics, England and Wales

A. Percentage Point Change in Proportion of Proven Offences Committed by Children, March 2014-2024



B. Number of Proven Offences by Children, by Offence Group and Gravity Score Band, 2024



Notes: The source for these statistics is the Youth Justice Statistics Report 2023-2024 [YJB 2024]: https://www.gov.uk/government/statistics/youth-justice-statistics-2023-to-2024/youth-justice-statistics-2023-to-2024. Panel A corresponds to Supplementary Tables Chapter 4, Table 4.1, and Panel B corresponds to Supplementary Tables Chapter 4, Table 4.6. In Panel B the time period refers to th eyear ending March 2025. An offence's gravity is scored out of 8, ranging from 1 (least serious) up to 8 (most serious).

Figure 2: Participant Flow and Timeline of Data Collection

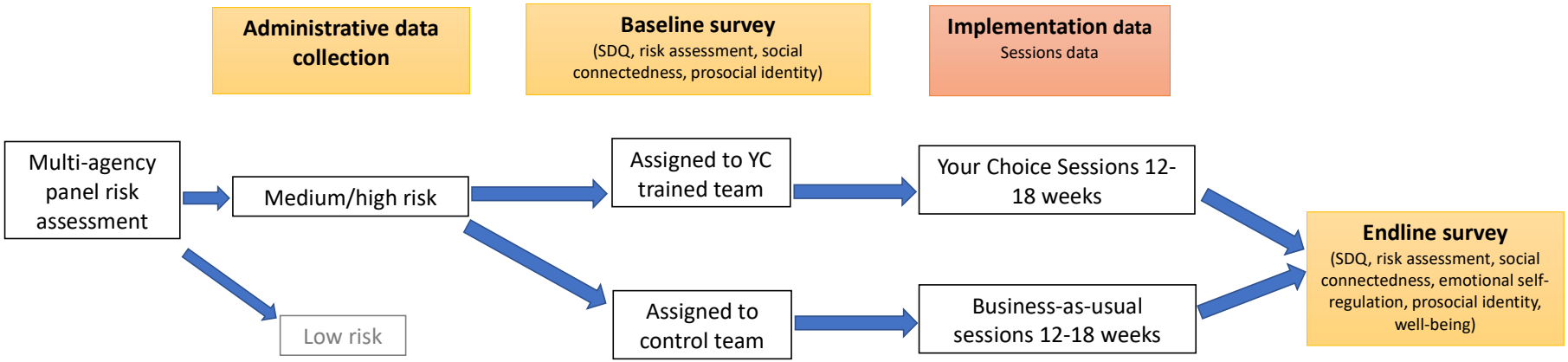
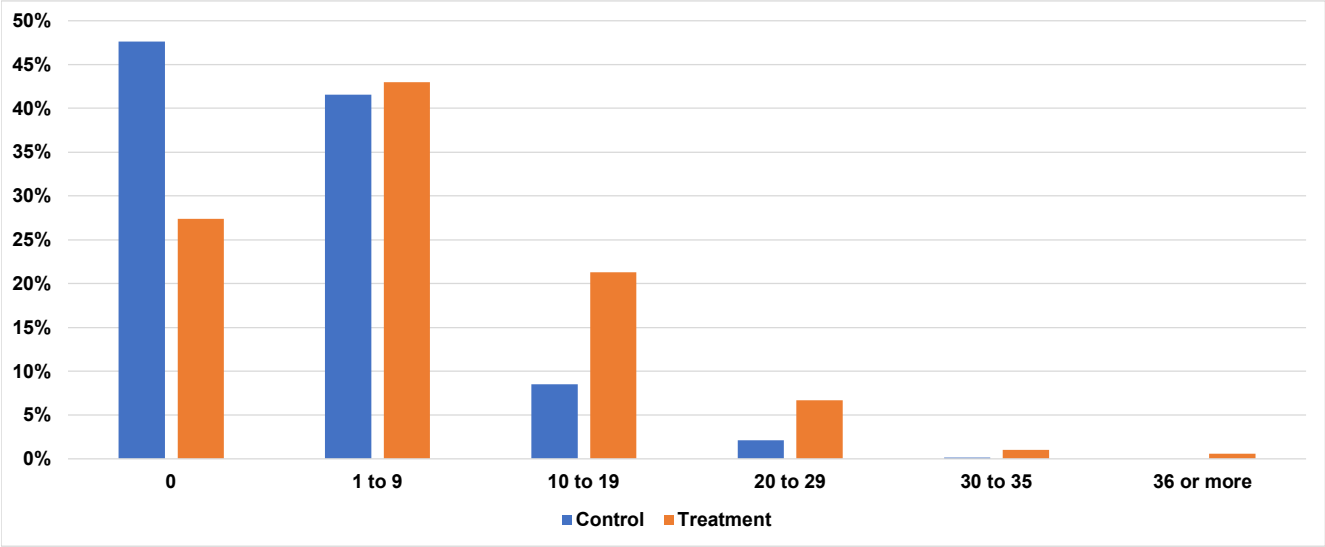
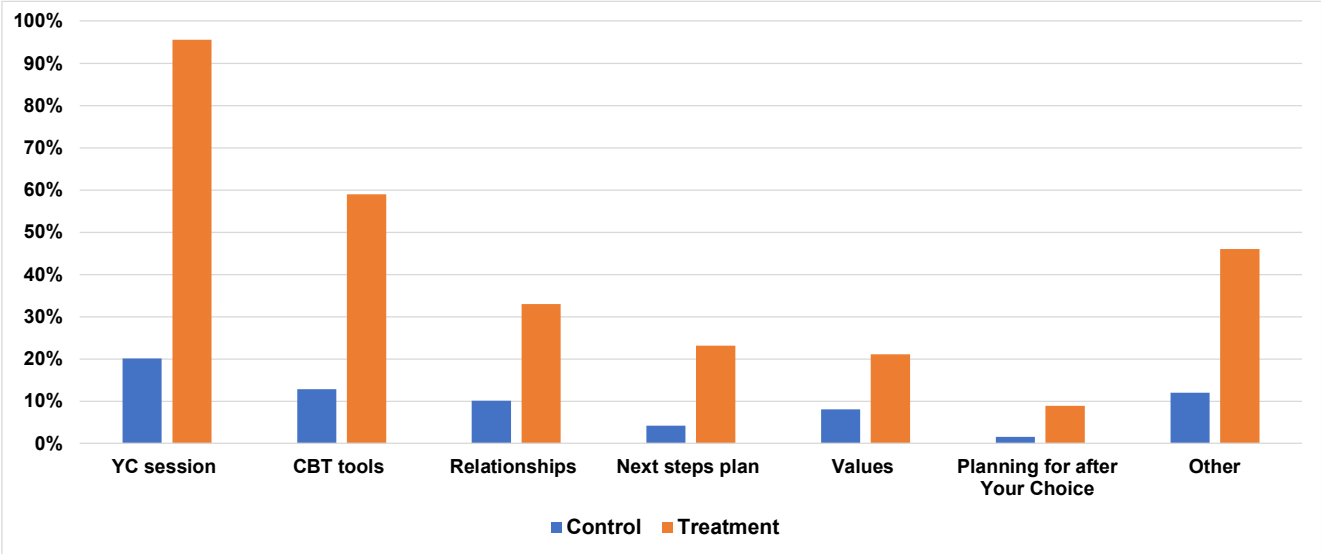


Figure 3: Young Person-Practitioner Sessions in the Scale-up

A. Attended Sessions



B. Session Content



Notes: All data are derived from the session forms collected from practitioners during the scale-up phase of the study. Panel B on session content shows the proportion of sessions in which a given content type was covered. Practitioners were asked: Please describe the content of the session by selecting as many options as apply from the list below. the 'other category' includes contact with siblings/parents, crisis, and other content type.

Table A1: Assignment of Young Persons to Services

Mean, p-values in brackets

	(1) Control	(2) Treatment	(3) p-value
A: Pilot			
Children Looked After	.658	.430	[.084]
Youth Justice	.228	.246	[.881]
Adolescent Safeguarding	.076	.123	[.481]
B: Scale-up			
Children Looked After	.080	.059	[.638]
Youth Justice	.082	.385	[.010]
Adolescent Safeguarding	.704	.406	[.043]
Schools/schools workers	.100	.107	[.927]

Notes: In Panel A, the sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. In Panel B, the sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. For each sample, we report means of the services they are assigned to in control and treated groups (Columns 1 and 2), and the p-value on a t-test of equality of means (Column 3). This is based on an OLS regression where standard errors are clustered at the Local Authority level, and in Panel B we also condition on local authority fixed effects.

Table A2: Baseline Balance on Subcomponents of Social Connectness Scale

Columns 1 to 4: Mean, standard deviation, raw difference and p-values in brackets
Columns 5 to 8: Mean, standard deviation, within LA differences and p-values in brackets

	A. Pilot				B. Scale-up			
	(1) Control	(2) Treatment	(3) Difference	(4) p-value	(5) Control	(6) Treatment	(7) Difference	(8) p-value
Number of observations	69	95			537	730		
Support from adult at home (1-5)	3.95	4.07	.113	[.441]	4.18	4.08	-.107	[.056]
	(1.02)	(.913)			(.835)	(.917)		
Support from adult at school (1-5)	3.33	3.58	.243	[.578]	3.76	3.66	-.098	[.360]
	(1.44)	(1.17)			(1.05)	(1.00)		
Support from adult elsewhere (1-5)	3.75	4.01	.259	[.196]	3.78	3.69	-.091	[.298]
	(1.17)	(1.07)			(1.13)	(1.12)		
Member of a club, sports team, church group, or other group (1-5)	2.32	2.03	-.292	[.256]	2.58	2.38	-.207	[.321]
	(1.58)	(1.41)			(1.63)	(1.55)		
Takes lessons in music, arts, sports or have a hobby (1-5)	2.56	2.52	-.038	[.870]	2.85	2.72	-.128	[.557]
	(1.60)	(1.51)			(1.62)	(1.57)		

Notes: In Panel A, the sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. In Panel B, the sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. All data is derived from the baseline survey to young persons. In Panel A we report means (standard deviations) of each characteristic in control and treated groups (Columns 1 and 2), their raw difference (Column 3), and the p-value on a t-test of equality of means (Column 4). This is based on an OLS regression where standard errors are clustered at the Local Authority level. In Panel B we report means in control and treated groups (Columns 5 and 6), their within local authority difference (Column 7), and the p-value on the corresponding t-test of equality of means (Column 8). This is based on an OLS regression where standard errors are clustered at the Local Authority level. The scores associated with ‘Support from adult at home’, ‘Support from adult at school’, and ‘Support from adult elsewhere’, `Member of a club, sports team, church group, or other group' and `Takes lessons in music, arts, sports or have a hobby' are scored on a 1-5 scale, where ‘None of the time’ is scored as 1, and ‘All the time’ is scored as 5.

Table A3: Attrition

Outcome: Adolescent tracked to endline
OLS estimates, standard errors clustered at the LA level

	Pilot			Scale-up			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Your Choice (treatment)	.113	.151	1.57***	-.108	-.015	-.054	-.144
	(.091)	(.106)	(.441)	(.064)	(.064)	(.047)	(.110)
Female		-.112	-.242*			-.028	.001
		(.104)	(.134)			(.032)	(.033)
Aged 16/17 (vs. aged 11-15)		-.055	-.113			-.056**	-.060
		(.138)	(.190)			(.027)	(.047)
White		.488*	.081			-.031	-.033
		(.253)	(.210)			(.039)	(.064)
Black		.283	-.302			-.028	-.062
		(.253)	(.271)			(.040)	(.058)
Asian		.355	-.291			.023	-.018
		(.294)	(.278)			(.032)	(.047)
Mixed race/ethnicity		.417*	-.336			-.016	-.054**
		(.238)	(.198)			(.028)	(.025)
Special education needs		-.062	-.375			.003	-.017
		(.071)	(.257)			(.035)	(.060)
Disability		.159	.139			.033	-.011
		(.115)	(.138)			(.031)	(.044)
In social care		-.088	.177			-.049	-.006
		(.141)	(.265)			(.040)	(.072)
Conduct problems: high/very high range		.010	.008			.016	-.012
		(.161)	(.202)			(.025)	(.051)
Self-assessed risk		-.042	-.120			-.018	-.023
		(.106)	(.120)			(.016)	(.020)
Practitioner-assessed risk		-.051	.497**			-.025	-.032**
		(.086)	(.176)			(.015)	(.014)
Control group mean	.430	.526	.526	.791	.791	.904	.904
Joint test on interactions [F-statistic]			[24.7]				[.783]
Strata FE (LA)	No	No	No	No	Yes	Yes	Yes
Characteristics	No	Yes	Yes	No	No	Yes	Yes
Your Choice x Characteristic interactions	No	No	Yes	No	No	No	Yes
Observations	193	94	94	1,465	1,465	1,051	1,051

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. In Columns 1 to 3, the sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. In Columns 4 to 7, the sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. The outcome is whether the young person was tracked to endline. OLS estimates are presented where standard errors are clustered at the Local Authority level. In Columns 3 and 7 the specification includes a set of interactions between the Your Choice treatment dummy and all individual characteristics. At the foot of Columns 3 and 7 we report the F-statistic on the joint null that all interactions are zero.

Table A4: Treatment Effects on Continuous SDQ Scores for Primary and Secondary Outcomes, Pilot

OLS estimates, standard errors clustered at the LA level, p-values in brackets
Outcomes: SDQ domain continuous scores

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors
	(1)	(2)	(3)	(4)	(5)
Your Choice	-.248	-.776	.159	-.123	.656*
	(.371)	(.566)	(.805)	(.720)	(.367)
p-val	*[.513]*	*[.189]*	*[.846]*	*[.867]*	*[.093]*
Baseline outcome	.600***	.709***	.615***	.485***	.598***
	(.085)	(.072)	(.073)	(.091)	(.056)
p-val	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*
Control Mean Baseline	3.70	5.23	4.00	3.12	7.77
Strata FE (LA)	No	No	No	No	No
Characteristics	No	No	No	No	No
Observations	81	80	79	78	78

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. OLS estimates are presented where standard errors are clustered at the Local Authority level, and p-values are reported in brackets. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score. Throughout we control only for the baseline outcome where available. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A5: Treatment Effects on Primary and Secondary Outcomes, Pilot Robustness

OLS estimates, standard errors clustered at the LA level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness	Emotional Self-regulation	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A. Controlling for Characteristics Imbalanced at Baseline										
Your Choice	-.105	-.084	.051	-.038	.001	.076	.014	.076	-.052	.149
	(.094)	(.102)	(.102)	(.148)	(.071)	(.147)	(.205)	(.204)	(.179)	(.188)
p-val	*[.280]*	*[.422]*	*[.624]*	*[.798]*	*[.993]*	*[.613]*	*[.946]*	*[.713]*	*[.777]*	*[.441]*
Control Mean Baseline	.333	.423	.308	.308	.154	3.81			2.28	2.89
Baseline outcome	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	81	80	79	78	78	78	96	85	78	80
B. LA Fixed Effects and Characteristics										
Your Choice	-.162	-.015	.048	-.121	.074	-.050	.098	-.089	-.093	.105
	(.129)	(.122)	(.135)	(.190)	(.088)	(.242)	(.333)	(.280)	(.248)	(.260)
p-val	*[.228]*	*[.901]*	*[.727]*	*[.533]*	*[.411]*	*[.838]*	*[.772]*	*[.753]*	*[.713]*	*[.690]*
Control Mean Baseline	.333	.423	.308	.308	.154	3.81			2.28	2.89
Baseline outcome	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	81	80	79	78	78	78	96	85	78	80

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. OLS estimates are presented where standard errors are clustered at the Local Authority level, and p-values are reported in brackets. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 9 and 10 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control for the baseline outcome where available. In Panel A we control for characteristics that are imbalanced at baseline (with p<.05): whether the young person is Asian, and whether their primary need is 'other'. In Panel B we control for these characteristics and local authority fixed effects. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A6: Treatment Effects on Primary and Secondary Outcomes, Pilot

OLS estimates, standard errors clustered at the team level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness	Emotional Self-regulation	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Your Choice	-.105	-.084	0.051	-0.038	.001	.076	.014	.076	-.052	.149
	(.083)	(.139)	(.092)	(.127)	(.076)	(.139)	(.195)	(.195)	(.142)	(.173)
p-val	*[.216]*	*[.552]*	*[.586]*	*[.765]*	*[.993]*	*[.587]*	*[.697]*	*[.943]*	*[.718]*	*[.396]*
Baseline outcome	.334***	.504***	.571***	.560***	.494***	.400***			.403***	.417***
	(.102)	(.099)	(.125)	(.092)	(.112)	(.100)			(.115)	(.098)
p-val	*[.003]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*			*[.002]*	*[.000]*
Control Mean Baseline	.333	.423	.308	.308	.154	3.81	-	-	2.28	2.89
Characteristics	No	No	No	No	No	No	No	No	No	No
Observations	81	80	79	78	78	78	96	85	78	80
Clusters (Teams)	29	29	28	28	29	29	33	30	29	30

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. OLS estimates are presented where standard errors are clustered at the team level, and p-values are reported in brackets. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 9 and 10 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A7: Treatment Effects on Primary and Secondary Outcomes, Pilot

OLS estimates, boostrapped standard errors clustered at the team level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness	Emotional Self-regulation	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Your Choice	-.105	-.084	.051	-.038	.001	.076	.014	.076	-.052	.149
	(.083)	(.148)	(.086)	(.122)	(.083)	(.147)	(.188)	(.195)	(.149)	(.189)
p-val	*[.205]*	*[.572]*	*[.555]*	*[.752]*	*[.994]*	*[.605]*	*[.684]*	*[.943]*	*[.730]*	*[.431]*
Baseline outcome	.334***	.504***	.571***	.560***	.494***	.400***			.403***	.417***
	(.106)	(.106)	(.131)	(.102)	(.107)	(.099)			(.124)	(.107)
p-val	*[.002]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*			*[.001]*	*[.000]*
Control Mean Baseline	.333	.423	.308	.308	.154	3.81	-	-	2.28	2.89
Characteristics	No	No	No	No	No	No	No	No	No	No
Observations	81	80	79	78	78	78	96	85	78	80
Clusters (Teams)	29	29	28	28	29	29	33	30	29	30

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. OLS estimates are presented where standard errors are bootstrapped and clustered at the team level, and p-values are reported in brackets. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 9 and 10 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A8: IPW Adjusted Primary and Secondary Outcomes, Pilot

OLS estimates, standard errors clustered at the local authority level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness	Emotional Self-regulation	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Your Choice	-.160	-.169	.054	-.193	-.042	.192	.394*	-.012	-.335	-.161
	(.123)	(.108)	(.108)	(.153)	(.094)	(.216)	(.208)	(.309)	(.235)	(.227)
p-val	*[.218]*	*[.144]*	*[.624]*	*[.232]*	*[.663]*	*[.390]*	*[.082]*	*[.969]*	*[.179]*	*[.492]*
Baseline outcome	.297**	.510***	.645***	.561***	0.392**	.126			.314	.336***
	(.119)	(.087)	(.091)	(.138)	(.171)	(.086)			(.218)	(.108)
p-val	*[.028]*	*[.000]*	*[.000]*	*[.002]*	*[.041]*	*[.169]*			*[.176]*	*[.009]*
Control Mean Baseline	.529	.588	.412	.353	.235	3.75			2.42	3.27
Characteristics	No	No	No	No	No	No	No	No	No	No
Observations	53	53	52	52	51	52	57	53	52	42

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. It is reweighted using inverse probability weights computed as the inverse of the predicted probabilities of being observed at endline, using demographic characteristics, baseline high/very high range of conduct problems, the self-reported *Checkpoint* risk score and the practitioner-assessed *Checkpoint* risk score at baseline as predictors of attrition. OLS estimates are presented where standard errors are clustered at the Local Authority level, and p-values are reported in brackets. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 9 and 10 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A9: Session Content in Line with YC Design

OLS estimates, standard errors clustered at the LA level, p-values in brackets
Outcome: Measure of CBT-oriented content

	Pilot		Scale-up	
	Young person	Practitioner	Young person	Practitioner
	(1)	(2)	(3)	(4)
Your Choice	-.135	.188	.170**	.243
	(.183)	(.140)	(.083)	(.167)
p-val	*[.470]*	*[.194]*	*[.050]*	*[.157]*
Control mean	4.18	4.04	3.99	3.80
Control SD	.727	.504	.859	1.103
LA Fixed Effects	No	No	Yes	Yes
Observations	86	101	1,030	1,150

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the pilot (Columns 1 and 2) and scale-up (Columns 3 and 4) phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. OLS estimates are presented where standard errors are clustered at the LA level and p-values are reported in brackets. The outcomes are averages of 7 questions, asked to the young person (Columns 1 and 3) and to the practitioner (Columns 2 and 4) measuring the extent to which the practitioner supported the young person with fidelity to the Your Choice intervention CBT-based training, on a 5 point Likert scale. At the foot of each column, we report the mean and standard deviation of the outcome in the control group.

Table A10: Conditional Logit Specification, Primary and Secondary Outcomes, Scale-up

Conditional logit estimates, standard errors clustered at the LA level, p-values in brackets

Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	(1) Conduct Problems	(2) Hyperactivity	(3) Emotional Problems	(4) Peer Problems	(5) Prosocial Behaviors
Your Choice	.170	-.044	-.069	-.168	.143
	(.176)	(.176)	(.184)	(.180)	(.228)
p-val	*[.335]*	*[.804]*	*[.707]*	*[.349]*	*[.529]*
Baseline outcome	1.49***	2.23***	2.12***	1.47***	1.75***
	(.246)	(.188)	(.188)	(.167)	(.214)
p-val	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*
Control Mean (SD) at Baseline	.235 (.431)	.304 (.460)	.200 (.400)	.281 (.450)	.179 (.384)
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes
Observations	933	958	917	936	953

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. Conditional logit estimates are presented where standard errors are clustered at the Local Authority level. The outcomes are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). Throughout we control only for the baseline outcome where available. We report conventional p-values. At the foot of each Column we report the mean (standard deviation) outcome among controls tracked to endline.

Table A11: Primary and Secondary Outcomes, Scale-up

OLS estimates, standard errors clustered at the LA level, p-value in brackets
Outcomes: SDQ measure scores

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors
	(1)	(2)	(3)	(4)	(5)
Your Choice	.027	.116	-.080	.027	-.139
	(.141)	(.200)	(.151)	(.150)	(.171)
p-val	*[.848]*	*[.568]*	*[.600]*	*[.858]*	*[.424]*
Baseline outcome	.471***	.608***	.572***	.476***	.558***
	(.030)	(.018)	(.019)	(.032)	(.031)
p-val	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*
Control Mean Baseline	3.03	4.96	3.17	2.64	7.29
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes
Characteristics	No	No	No	No	No
Observations	962	962	957	963	967

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. OLS estimates are presented where standard errors are clustered at the Local Authority level. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score. Throughout we control only for the baseline outcome where available. We report conventional p-values. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A12: Treatment Effects on Primary and Secondary Outcomes, Scale-up Robustness

OLS estimates, standard errors clustered at the LA level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness scale	Emotional Self-regulation	Prosocial identity scale	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Your Choice	.000	-.036	-.016	-.041	.002	.094**	.070	.050	.016	-.073	.160
	(.029)	(.033)	(.020)	(.035)	(.043)	(.045)	(.065)	(.061)	(.068)	(.067)	(.161)
p-val	*[.994]*	*[.276]*	*[.436]*	*[.253]*	*[.955]*	*[.043]*	*[.285]*	*[.418]*	*[.820]*	*[.288]*	*[.326]*
adj p-val		*[.739]*	*[.739]*	*[.739]*	*[.944]*	*[.739]*	*[.739]*	*[.739]*	*[.942]*	*[.739]*	*[.739]*
Baseline outcome	.244***	.428***	.374***	.314***	.130**	.436***		.098*		.278***	.530***
	(.044)	(.035)	(.037)	(.037)	(.060)	(.035)		(.050)		(.062)	(.065)
p-val	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.039]*	*[.000]*		*[.063]*		*[.000]*	*[.000]*
Control Mean Baseline	.243	.307	.207	.276	.180	3.73		2.85		1.63	2.29
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	878	878	875	879	897	900	929	892	894	871	936

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. OLS estimates are presented where standard errors are clustered at the Local Authority level. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 prosocial identity is measured by the Prosocial Identity Scale. In Column 9 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 10 and 11 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available. We also control for characteristics that are imbalanced at baseline (with p<.05): whether the young person reports having special education needs, their prosocial behaviors score (0-10), and their practitioner-assessed Checkpoint aggregate score. We report conventional p-values and p-values that adjust for multiple hypothesis testing (MHT). These account for MHT across Columns 2 to 5, 6 to 9, and 10 and 11. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A13: Treatment Effects on Primary and Secondary Outcomes, Scale-up Post-Double Lasso

OLS estimates, standard errors clustered at the LA level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness scale	Emotional Self-regulation	Prosocial identity scale	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Your Choice	.014	-.013	-.015	-.046	.016	.046	.076	.021	.015	-.071	.160
	(.027)	(.028)	(.019)	(.032)	(.035)	(.048)	(.054)	(.040)	(.066)	(.060)	(.122)
p-val	*[.614]*	*[.650]*	*[.425]*	*[.146]*	*[.643]*	*[.345]*	*[.158]*	*[.601]*	*[.821]*	*[.240]*	*[.188]*
Control Mean Baseline	.177	.268	.181	.324	.223	3.75	1.64	3.67	3.54	1.54	1.76
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	962	962	957	963	967	994	1,066	983	1,024	961	1,091

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. Post-Double Selection Lasso estimates are presented where standard errors are clustered at the Local Authority level. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 prosocial identity is measured by the Prosocial Identity Scale. In Column 9 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 10 and 11 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control for the baseline outcome where available and baseline control variables selected to best predict the outcome and the treatment via Post-Double Selection Lasso. At the foot of each Column we report the mean outcome among controls tracked to endline and the selected control variables. The sample size is slightly different from the main analytical sample size because we impute missing baseline variables to maximise the sample to run the post double selection Lasso.

Table A14: Treatment Effects on Primary and Secondary Outcomes, Scale-up

OLS estimates, standard errors clustered at the team level, p-values in brackets
Outcomes Columns 1 to 4: ln high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness scale	Emotional Self-regulation	Prosocial identity scale	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Your Choice	.028	-.010	-.007	-.032	.025	.046	-.017	-.025	.056	-.062	.199*
	(.028)	(.029)	(.023)	(.037)	(.030)	(.049)	(.051)	(.062)	(.054)	(.059)	(.106)
p-val (unadjusted)	*[.315]*	*[.722]*	*[.758]*	*[.381]*	*[.394]*	*[.354]*	*[.745]*	*[.693]*	*[.307]*	*[.289]*	*[.064]*
adj p-val (LSV)		*[.939]*	*[.793]*	*[.882]*	*[.827]*	*[.873]*	*[.806]*	*[.784]*	*[.936]*	*[.378]*	*[.359]*
Baseline outcome	.264***	.453***	.370***	.309***	.356***	.460***	.190***			.319***	.590***
	(.042)	(.033)	(.035)	(.035)	(.040)	(.034)	(.047)			(.044)	(.088)
p-val	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*			*[.000]*	*[.000]*
Control Mean Baseline	.246	.304	.199	.281	.178	3.73		2.85		1.63	2.30
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	962	962	957	963	967	994	1,066	983	1,024	961	1,091
Clusters	109	108	109	109	109	111	112	111	111	110	112

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. OLS estimates are presented where standard errors are clustered at the team level. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 prosocial identity is measured by the Prosocial Identity Scale. In Column 9 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 10 and 11 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available. We report conventional p-values, p-values that adjust for multiple hypothesis testing using List, Shaikh and Vayalinka (2021)'s approach (LSV). These account for MHT using LSV (2021) across Columns 2 to 5, 6 to 9, and 10 and 11. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A15: Treatment Effects on Primary and Secondary Outcomes, Scale-up

OLS estimates, bootstrapped standard errors clustered at the team level, p-values in brackets
Outcomes Columns 1 to 4: ln high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness scale	Emotional Self-regulation	Prosocial identity scale	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Your Choice	.028	-.010	-.007	-.032	.025	.046	.056	-.017	-.025	-.062	.199*
	(.039)	(.036)	(.014)	(.061)	(.049)	(.094)	(.059)	(.065)	(.076)	(.064)	(.119)
p-val (unadjusted)	*[.471]*	*[.776]*	*[.598]*	*[.597]*	*[.605]*	*[.625]*	*[.779]*	*[.703]*	*[.462]*	*[.331]*	*[.096]*
adj p-val (LSV)		*[.939]*	*[.793]*	*[.882]*	*[.827]*	*[.873]*	*[.806]*	*[.784]*	*[.936]*	*[.378]*	*[.359]*
Baseline outcome	.264***	.453***	.370***	.309***	.356***	.460***		.190***		.319***	.590***
	(.041)	(.029)	(.025)	(.014)	(.049)	(.030)		(.039)		(.051)	(.072)
p-val	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*		*[.000]*		*[.000]*	*[.000]*
Control Mean Baseline	.246	.304	.199	.281	.178	3.73		2.85		1.63	2.30
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	962	962	957	963	967	994	1,066	983	1,024	961	1,091
Clusters	109	108	109	109	109	111	112	111	111	110	112

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. OLS estimates are presented where standard errors are bootstrapped and clustered at the team level. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 prosocial identity is measured by the Prosocial Identity Scale. In Column 9 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 10 and 11 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available. We report conventional p-values, p-values that adjust for multiple hypothesis testing using Romano Wolf algorithm (RW), and p-values that adjust for multiple hypothesis testing using List Shaikh Vayalinka (2021) approach (LSV). In Columns 10 and 11, the conventional (analytical) p-value is higher than the adjusted one, but the unadjusted bootstrapped one (not reported) is, as expected, lower than the adjusted (bootstrapped) one. These account for MHT using Romano Wolf across Columns 2 to 5, 6 to 9, and 10 and 11. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A16: IPW Adjusted Primary and Secondary Outcomes, Scale-up

OLS estimates, standard errors clustered at the local authority level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness scale	Emotional Self-regulation	Prosocial identity scale	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Your Choice	.020	-.018	-.006	-.056	.018	.085*	-.003	-.032	.114*	-.045	.092
	(.027)	(.033)	(.020)	(.037)	(.046)	(.046)	(.060)	(.065)	(.059)	(.065)	(.154)
p-val	*[.470]*	*[.590]*	*[.768]*	*[.147]*	*[.701]*	*[.074]*	*[.958]*	*[.620]*	*[.064]*	*[.492]*	*[.556]*
p-val (RW)		*[.929]*	*[.737]*	*[.771]*	*[.853]*	*[.066]*	*[.066]*	*[.940]*	*[.745]*	*[.445]*	*[.445]*
Baseline outcome	.271***	.448***	.376***	.309***	.327***	.444***	.177***			.320***	.506***
	(.045)	(.034)	(.044)	(.037)	(.047)	(.039)	(.060)			(.051)	(.065)
p-val	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.000]*	*[.004]*			*[.000]*	*[.000]*
Control Mean Baseline	.237	.302	.205	.267	.174	3.75	2.87			1.63	2.29
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	842	830	829	831	838	851	844	846	877	832	825

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. OLS estimates are presented where standard errors are clustered at the Local Authority level. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 prosocial identity is measured by the Prosocial Identity Scale. In Column 9 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 10 and 11 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available. We report conventional p-values, and p-values that adjust for multiple hypothesis testing using Romano Wolf algorithm (RW). We use the RW method over the LSV method because we cannot apply reweighting with the LSV command in Stata. In Columns 10 and 11, the conventional (analytical) p-value is higher than the adjusted one, but the unadjusted bootstrapped one (not reported) is, as expected, lower than the adjusted (bootstrapped) one. These account for MHT across Columns 2 to 5, 6 to 9, and 10 and 11. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A17: Child Services, Scale-up Robustness

OLS estimates, standard errors clustered at the LA level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness scale	Emotional Self-regulation	Prosocial identity scale	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Your Choice	.042	-.009	-.008	-.022	.039	.057	.106*	-.044	-.091	-.075	.126
	(.030)	(.036)	(.023)	(.040)	(.043)	(.056)	(.060)	(.059)	(.060)	(.072)	(.100)
p-val	*[.167]*	*[.795]*	*[.726]*	*[.576]*	*[.367]*	*[.322]*	*[.089]*	*[.463]*	*[.142]*	*[.308]*	*[.217]*
adj p-val		*[.876]*	*[.876]*	*[.844]*	*[.611]*	*[.337]*	*[.084]*	*[.337]*	*[.130]*	*[.122]*	*[.104]*
Control Mean Baseline	.246	.304	.199	.281	.178	3.73		2.85		1.63	2.30
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Services LA	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	962	962	957	963	967	994	1,066	983	1,024	961	1,091

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. OLS estimates are presented where standard errors are clustered at the Local Authority level. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 prosocial identity is measured by the Prosocial Identity Scale. In Column 9 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 10 and 11 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available. We also control for characteristics that are imbalanced at baseline (with p<.05): whether the young person reports having special education needs, their prosocial behaviors score (0-10), and their practitioner-assessed Checkpoint aggregate score. We report conventional p-values and p-values that adjust for multiple hypothesis testing (MHT). These account for MHT across Columns 2 to 5, 6 to 9, and 10 and 11. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A18: Background Characteristics at Baseline, Pilot vs Scale-up

Mean, difference and p-values in brackets

	(1) Pilot	(2) Scale-up	(3) Difference	(4) Normalized Difference	(5) p-value
Number of observations	193	1340			
A: Demographics					
Female	.303	.324	-.021	-.046	[.606]
Aged 16/17 (vs. aged 11-15)	.466	.497	-.031	-.061	[.778]
White	.244	.192	.052	.130	[.443]
Black	.254	.217	.037	.088	[.454]
Asian	.083	.122	-.039	-.123	[.372]
Mixed race/ethnicity	.145	.254	-.109	-.256	[.007]
Special education needs	.262	.190	.072	.181	[.136]
Disability	.471	.264	.206	.457	[.020]
In social care	.161	.148	.013	.037	[.762]
B: Primary Need Assessment					
Family dysfunction	.197	.078	.119	.413	[.029]
Socially unacceptable behavior	.212	.382	-.170	-.353	[.029]
Family in acute stress	.073	.060	.012	.052	[.616]
Other	.192	.264	-.072	-.166	[.376]
Not stated	.326	.216	.111	.264	[.227]
C: Activities					
School/college	.689	.741	-.052	-.118	[.486]
Work	.036	.048	-.012	-.058	[.582]
Training	.048	.039	.009	.045	[.720]
Other	.228	.194	.033	.083	[.530]

Notes: The sample are those young people recruited in the pilot phase and in the scale-up phase of the studdy and that were assigned to treatment or control among practitioner teams randomized into their status. We report means of each characteristic in the pilot and efficacy samples (Columns 1 and 2), their raw difference (Column 3), their raw difference standardised by the standard deviation of the variable in the pooled sample (Column 4), and the p-value on a t-test of equality of means (Column 5). This is based on an OLS regression where standard errors are clustered at the Local Authority level. Panels A and B are derived from Local Authority Workbooks. Panel C is derived from the baseline survey to young persons.

Table A19: Background Characteristics at Baseline, Pilot vs Scale-up

Mean, standard deviation, difference and p-values in brackets

	(1) Pilot	(2) Scale-up	(3) Difference	(4) Normalized Difference	(5) p-value
A. Primary Outcome					
Conduct problems score (0-10)	3.74	3.41	.328	.148	[.364]
	(2.21)	(2.21)			
Conduct problems: high/very high range	.389	.304	.085	.183	[.250]
B. Secondary Outcomes					
Hyperactivity score (0-10)	5.64	5.34	.305	.115	[.475]
	(2.66)	(2.65)			
Hyperactivity score: high/very high range	.400	.350	.050	.104	[.489]
Emotional problems score (0-10)	3.50	3.10	.408	.173	[.221]
	(2.53)	(2.33)			
Emotional problems: high/very high range	.217	.169	.048	.126	[.308]
Peer problems score (0-10)	2.76	2.89	-.122	-.066	[.614]
	(1.95)	(1.82)			
Peer problems: high/very high range	.342	.327	.014	.031	[.791]
Prosocial behaviors score (0-10)	6.84	7.03	-.187	-.094	[.267]
	(2.12)	(1.98)			
Prosocial behaviors: low/very low range	.252	.220	.031	.075	[.430]
Social connectedness scale (1-5)	3.62	3.66	-.040	-.054	[.622]
	(0.79)	(0.75)			

Notes: The sample are those young people recruited in the pilot phase and in the scale-up phase of the studdy and that were assigned to treatment or control among practitioner teams randomized into their status. We report means of each characteristic in the pilot and efficacy samples (Columns 1 and 2), their raw difference (Column 3), their raw difference standardised by the standard deviation of the variable in the pooled sample (Column 4), and the p-value on a t-test of equality of means (Column 5). This is based on an OLS regression where standard errors are clustered at the Local Authority level.. This is based on an OLS regression where standard errors are clustered at the Local Authority level. Panels A and B are derived from the baseline survey to young persons. The outcomes are subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). The Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere.

Table A20: Background Characteristics at Baseline, Pilot vs Scale-up

Mean, difference and p-values in brackets

	(1) Pilot	(2) Scale-up	(3) Difference	(4) Normalized Difference	(5) p-value
A: Self-Assessed Risks					
Checkpoint safety score (1-5)	2.04	1.78	.257	.310	[.018]
	(.816)	(.827)			
In the next month, likely/very likely to:					
to cause damage in public place	.098	.076	.023	.084	[.425]
to get in trouble with police	.104	.069	.035	.133	[.234]
to physically hurt someone	.088	.076	.012	.046	[.651]
others will involve in any of above	.130	.122	.007	.022	[.775]
to witness a violent crime	.171	.183	-.012	-.031	[.632]
to be victim of violent crime	.047	.068	-.021	-.087	[.185]
B: Practitioner Assessed Risks					
Checkpoint safety score (1-5)	3.01	2.59	.420	.484	[.000]
	(.791)	(.866)			
In the next month, likely/very likely to:					
to cause damage in public place	.218	.173	.045	.117	[.191]
to get in trouble with police	.306	.245	.061	.139	[.276]
to physically hurt someone	.176	.170	.007	.017	[.819]
others will involve in any of above	.440	.383	.058	.118	[.358]
to witness a violent crime	.316	.209	.107	.257	[.041]
to be victim of violent crime	.176	.132	.045	.130	[.273]

Notes: The sample are those young people recruited in the pilot phase and in the scale-up phase of the studdy and that were assigned to treatment or control among practitioner teams randomized into their status. We report means of each characteristic in the pilot and efficacy samples (Columns 1 and 2), their raw difference (Column 3), their raw difference standardised by the standard deviation of the variable in the pooled sample (Column 4), and the p-value on a t-test of equality of means (Column 5). This is based on an OLS regression where standard errors are clustered at the Local Authority level. This is based on an OLS regression where standard errors are clustered at the Local Authority level. Panel A is derived from the Checkpoint risk assessment tool administered as part of the baseline survey to young persons. Panel B is derived from the Checkpoint risk assessment tool administered as part of baseline survey to practitioners.

Table A21: Sample Unrepresentativeness, All Outcomes

OLS estimates, standard errors clustered at the LA level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness scale	Emotional Self-regulation	Prosocial identity scale	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
A. Pilot											
Your Choice	-.105	-.084	.051	-.038	.001	.076	.014		.076	-.052	.149
	(.094)	(.102)	(.102)	(.148)	(.071)	(.147)	(.205)		(.204)	(.179)	(.188)
p-val	*[.280]*	*[.422]*	*[.624]*	*[.798]*	*[.993]*	*[.613]*	*[.946]*		*[.713]*	*[.777]*	*[.441]*
B. Entropy Balanced, LASSO Selected Scale-up											
Your Choice	-.106	*.005*	.374***	-.031	-.093	.083	.253	-.074	.371***	-.303	.212
	(.152)	(.134)	(.130)	(.093)	(.103)	(.156)	(.187)	(.325)	(.104)	(.278)	(.479)
p-val	*[.492]*	*[.973]*	*[.008]*	*[.738]*	*[.372]*	*[.599]*	*[.187]*	*[.821]*	*[.001]*	*[.284]*	*[.662]*
Control Mean Baseline	.246	.304	.199	.281	.178	3.73		2.85		1.63	2.30
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	843	832	840	834	839	851	879	844	846	824	889

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. In Panel B, the scale-up sample is reweighted using entropy weighting such that the first moments of a set of variables in the reweighted scale-up sample match those of the pilot sample. This set of variables is selected, among all baseline variables, by LASSO as the best predictors of being in the pilot sample and includes indicators for white, black, special education needs, disability, in social care, the primary needs assessment being family dysfunction, the primary needs assessment being being family stress, the primary needs assessment being not stated, baseline conduct problems in the high/very high range, baseline emotional problems in the high/very high range, the practitioner-assessed *Checkpoint* safety score, and the service assigned to -- Children Looked After, Early Help, and Youth Justice. OLS estimates are presented where standard errors are clustered at the Local Authority level. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 prosocial identity is measured by the Prosocial Identity Scale. In Column 9 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 10 and 11 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available. We report conventional p-values and p-values that adjust for multiple hypothesis testing (MHT). These account for MHT across Columns 2 to 5, 6 to 9, and 10 and 11. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A22: Correlations in SDQ Domains

A. Pilot sample

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors
Conduct Problems	1				
Hyperactivity	.546	1			
Emotional Problems	.367	.431	1		
Peer Problems	.166	.215	.401	1	
Prosocial Behaviors	-.298	-.272	-.064	-.245	1

B. MCS Age 14 Sample

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors
Conduct Problems	1				
Hyperactivity	.546	1			
Emotional Problems	.367	.350	1		
Peer Problems	.361	.342	.490	1	
Prosocial Behaviors	-.462	-.369	-.190	-.297	1

Notes: In Panel A, the sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. In Panel B, the sample are from the MCS age 14 data. The outcomes are subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score.

Table A23: IV Estimates, First Stage

OLS estimates, standard errors clustered at the LA level, p-value in brackets

	(1) At least one YC session	(2) Number of YC Sessions
Assigned to a Your Choice team	.516***	7.72***
	(.063)	(1.40)
	[.000]	*[.000]*
Control mean (sd)	.209 (.407)	.721 (2.31)
Strata FE (LA)	Yes	Yes
Characteristics	Yes	Yes
Observations	1,247	1,247

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams.OLS estimates are presented where standard errors are clustered at the Local Authority level. In both columns the outcome is an indicator for being assigned to a treatment team. In Column 1 the independent variable is an indicator for the young person receiving at least one Your Choice session. In Column 2 the independent variable is the number of Your Choice sessions received. The characteristic we control for is an indicator for individual scores on the Conduct problem dimension of the Strengths and Difficulties Questionnaire in the abnormal range. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table A24: Dosage Estimates at Scale-up

OLS estimates, standard errors clustered at the LA level, p-values in brackets
Outcomes Columns 1 to 4: In high/very high range of SDQ domain (or low/very low range in Column 5)

	Conduct Problems	Hyperactivity	Emotional Problems	Peer Problems	Prosocial Behaviors	Social connectedness scale	Emotional Self-regulation	Prosocial identity scale	Wellbeing	Checkpoint: Self-Assessed	Checkpoint: Practitioner-Assessed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
A. At least one session	.049	-.018	-.012	-.056	.044	.079	.099	-.029	-.042	-.108	.387
	(.050)	(.050)	(.035)	(.055)	(.063)	(.086)	(.121)	(.083)	(.125)	(.102)	(.246)
p-val	*[.328]*	*[.722]*	*[.726]*	*[.312]*	*[.487]*	*[.355]*	*[.414]*	*[.732]*	*[.734]*	*[.290]*	*[.116]*
B. Number of sessions	.003	-.001	-.001	-.004	.003	.005	.006	-.002	-.003	-.007	.025*
	(.003)	(.003)	(.002)	(.003)	(.004)	(.005)	(.008)	(.005)	(.008)	(.006)	(.015)
p-val	*[.364]*	*[.723]*	*[.719]*	*[.285]*	*[.504]*	*[.329]*	*[.430]*	*[.734]*	*[.741]*	*[.288]*	*[.091]*
Control Mean Baseline	.246	.304	.199	.281	.178	3.73		2.85		1.63	2.30
Strata FE (LA)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes
Observations	962	962	957	963	967	994	1066	983	1024	961	1091

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. The sample are those young people recruited in the scale-up phase of the study and that were assigned to treatment or control practitioner teams. Second stage IV estimates are presented where standard errors are clustered at the Local Authority level. The outcomes in Columns 1 to 5 are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score, and the indicator on any of the subscales indicates the individual scores in the high/very high range (except for the prosocial subscale where this indicates scoring low/very low). In Column 6 the Social Connectedness Scale measures the young person's resilience through strengthened sources of social support at home, school and elsewhere. In Column 7 emotional self-regulation is measured by the Self-regulation subscale from the Trait Emotional Intelligence Questionnaire -- Adolescent Short Form. In Column 8 prosocial identity is measured by the Prosocial Identity Scale. In Column 9 wellbeing is measured by the Short Warwick--Edinburgh Mental Well-being Scale. Columns 10 and 11 report the aggregate score from the Checkpoint risk assessment tool administered to young persons and practitioners. Throughout we control only for the baseline outcome where available. We report conventional p-values. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table C1: Strength and Difficulties Questionnaire

	Question Wording	Question Scoring		
		Not true	Somewhat true	Certainly true
Conduct problems				
Q1	I get very angry and often lose my temper	0	1	2
Q2	I usually do as I am told	2	1	0
Q3	I fight a lot. I can make other people do what I want	0	1	2
Q4	I am often accused of lying or cheating	0	1	2
Q5	I take things that are not mine from home, school or elsewhere	0	1	2
Hyperactivity				
Q1	I am restless, I cannot stay still for long	0	1	2
Q2	I am constantly fidgeting or squirming	0	1	2
Q3	I am easily distracted, I find it difficult to concentrate	0	1	2
Q4	I think before I do things	2	1	0
Q5	I finish the work I'm doing. My attention is good	2	1	0
Emotional problems				
Q1	I get a lot of headaches, stomach-aches or sickness	0	1	2
Q2	I worry a lot	0	1	2
Q3	I am often unhappy, down-hearted or tearful	0	1	2
Q4	I am nervous in new situations. I easily lose confidence	0	1	2
Q5	I have many fears, I am easily scared	0	1	2
Peer problems				
Q1	I am usually on my own. I generally play alone or keep to myself	0	1	2
Q2	I have one good friend or more	2	1	0
Q3	Other people my age generally like me	2	1	0
Q4	Other children or young people pick on me or bully me	0	1	2
Q5	I get on better with adults than with people my own age	0	1	2
Prosocial behaviors				
Q1	I try to be nice to other people. I care about their feelings	0	1	2
Q2	I usually share with others (food, games, pens etc.)	0	1	2
Q3	I am helpful if someone is hurt, upset or feeling ill	0	1	2
Q4	I am kind to younger children	0	1	2
Q5	I often volunteer to help others (parents, teachers, children)	0	1	2

Table C2: Checkpoint Scale [Goodman *et al*. 2022]

In this set of questions, we would like to ask you **how likely you think different situations will happen in the next month**.

Your responses are confidential – we will not tell anyone (e.g. your youth worker, the police) what you tell us. The only exception is if there is a risk of significant harm to you or other people. For questions marked with ** if you answer likely or very likely then we will let your practitioner know so that they can work with you towards a plan to keep you safe. If you have questions about this, please pause to talk to your practitioner.

In the next month, how likely do you think you are to engage in behaviour that…

	Very unlikely	**Unlikely**	**Neither likely nor unlikely**	**Likely**	**Very likely**
… could cause distress or damage in a public place? (e.g. graffiti or other damage to public property, being drunk in a public place, transport fare dodging)					
… would get you into trouble with the police? (e.g. theft, fire setting, selling illegal substances or property, drink or drug driving, taking a car without consent, carrying an item that could cause serious physical harm to others)					
… could physically hurt other people? ** (e.g. serious physical fights, using an item that could cause serious physical harm to others)					

In the next month, how likely do you think it is that…

	Very unlikely	**Unlikely**	**Neither likely nor unlikely**	**Likely**	**Very likely**
… that other people in your life will try and involve you in any of the above behaviours? **					

In the next month, how likely do you think it is that you will …

	Very unlikely	**Unlikely**	**Neither likely nor unlikely**	**Likely**	**Very likely**
… witness a violent crime?					
… be a victim of a violent crime? **					

Table C3: Intervention Fidelity Measures

Instrument in the Endline Young Person Questionnaire

Thinking about the worker you've spent most time with over the past 3-4 months, how often did they do the following things?

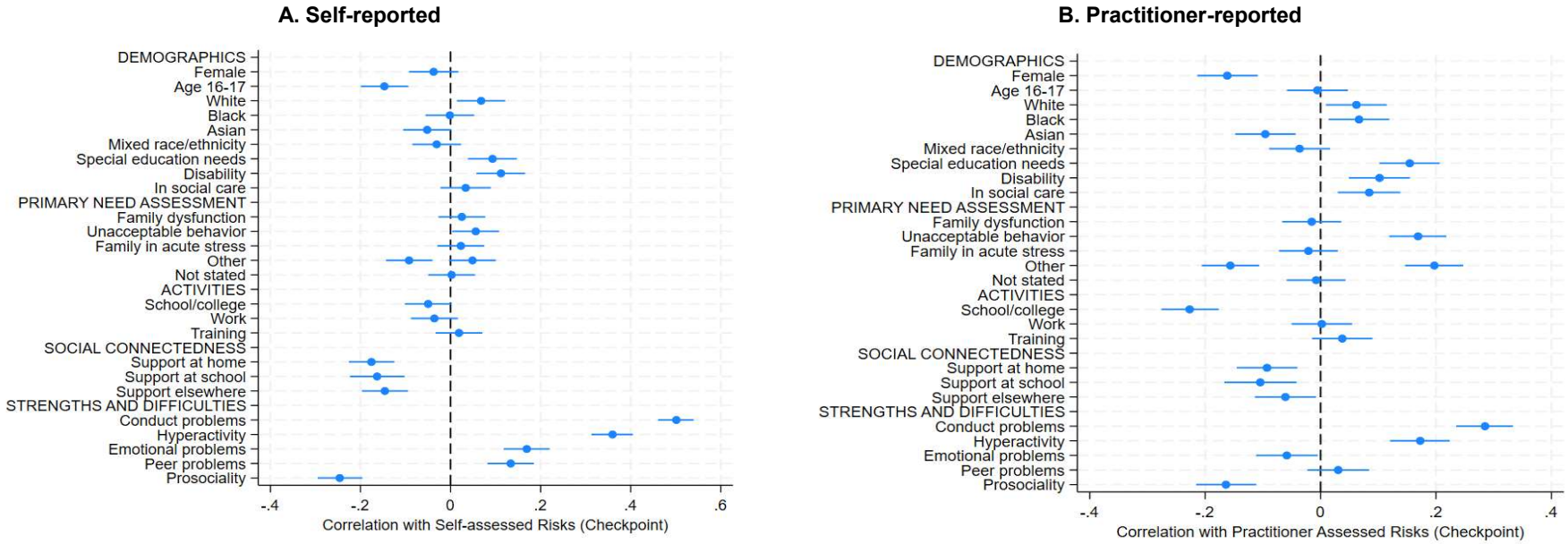
	Never	Rarely	Sometimes	Often	Always
My practitioner spent time getting to know and understand me.					
My practitioner tried to help me to better understand things that are important to me.					
My practitioner encouraged me to work towards goals that are important to me.					
My practitioner taught me new skills or behaviours.					
My practitioner helped me to see things differently.					
My practitioner helped me to find new ways of coping with difficult feelings.					
My practitioner helped me to find new ways of coping with difficult situations.					

Instrument in the Endline Practitioner Questionnaire

Thinking about the sessions you had with the young person over the past 3 - 4 months, how often did you do the following things?

	Never	Rarely	Sometimes	Often	Always
I spent time getting to know and understand the young person.					
I tried to help the young person to better understand things that are important to them.					
I encouraged the young person to work towards goals that are important to them.					
I taught the young person new skills or behaviours.					
I helped the young person to see things differently.					
I helped the young person to find new ways of coping with difficult feelings.					
I helped the young person to find new ways of coping with difficult situations.					

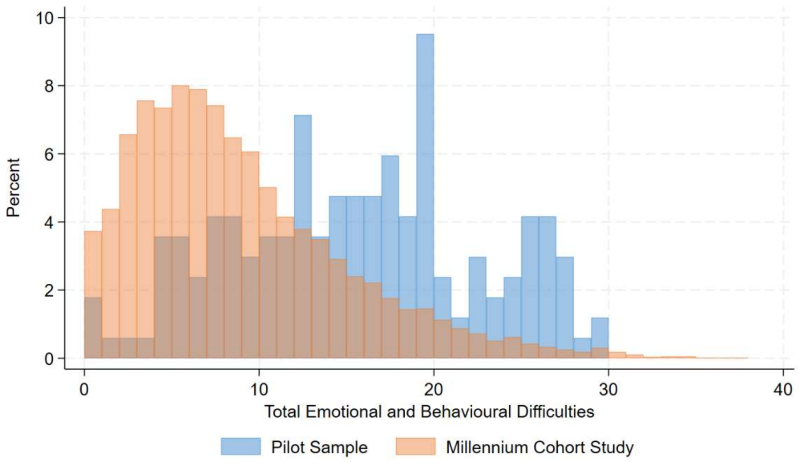
Figure A1: External Validity of the Checkpoint Scale



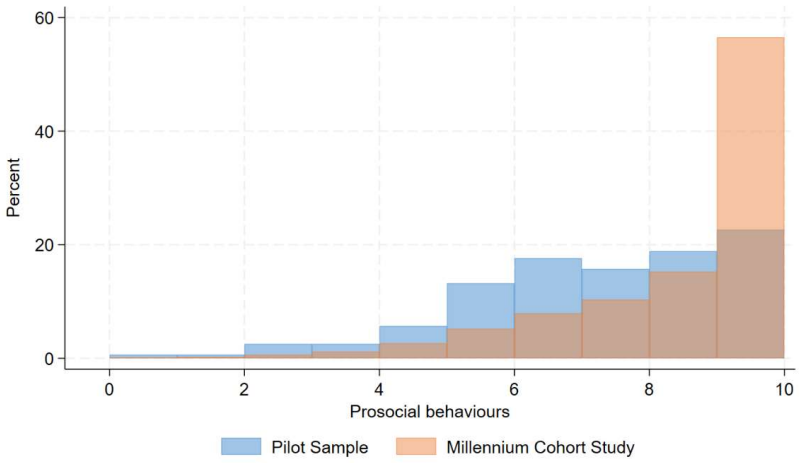
Notes: The sample are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status. Panel A is derived from the Checkpoint risk assessment tool administered as part of the baseline survey to young persons. Panel B is derived from the Checkpoint risk assessment tool administered as part of baseline survey to practitioners.

Figure A2: Comparison of the SDQ Measures to the MCS

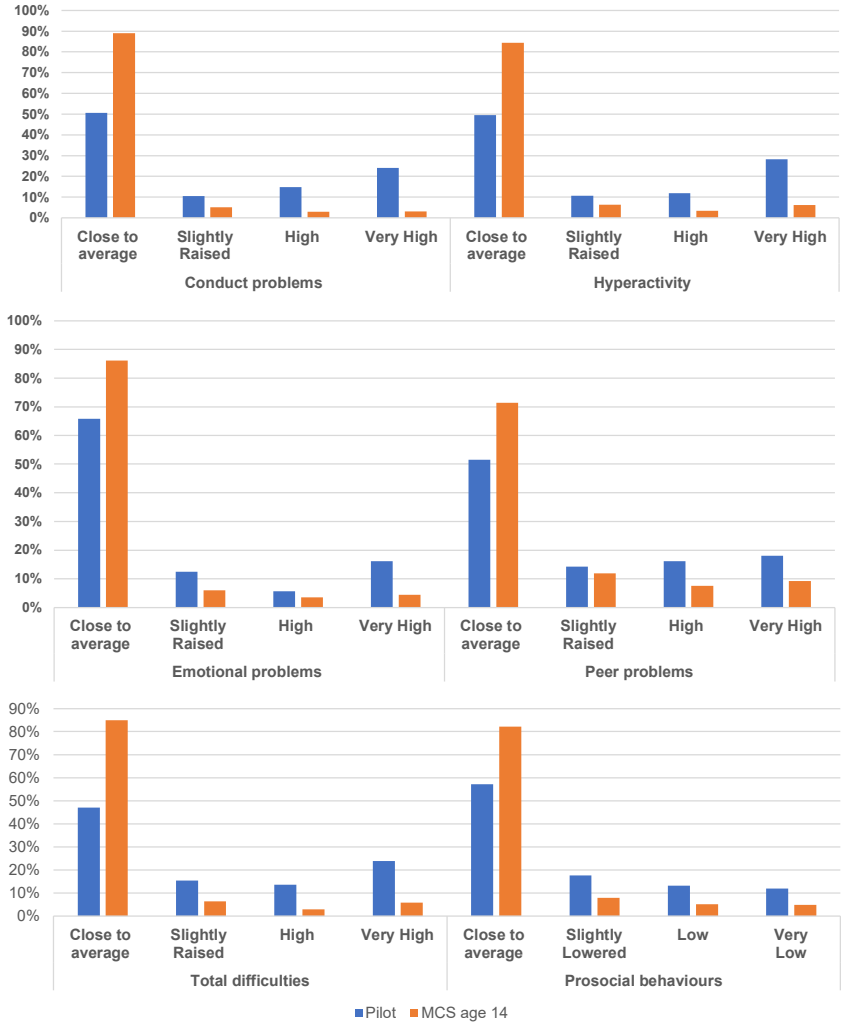
A. Distribution of SDQ Total Difficulties Score



B. Distribution of SDQ Prosocial Score



C. Proportion of Children in Close to Average, Slightly Raised (Lowered), High (Low), and Very High (Very Low) range



Notes: In Panels A and B, the samples are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status, and the from the MCS age 14 data. In Panel C the samples are those young people recruited in the pilot phase of the study and that were assigned to treatment or control among practitioner teams randomized into their status, the MCS age 14 data, and a benchmark set of population data. The outcomes are based on subscales from the Strengths and Difficulties Questionnaire: conduct problems, hyperactivity, emotional problems, peer problems, and prosocial behavior. Each subscale is measured on a ten-item score. The categorisation of scores into "Close to average", "Slightly Raised/Lowered", "High/Low" and "Very High/Very Low" followed the newer 4-band categorisation provided by the SDQ team.