



An assessment of behavioural change associated with the introduction of the Teaching Family Model to a group home residential placement

The planned introduction of the Teaching Family Homes (TFH) model of care to the Te Whare Tumanako o Kia Puawai residence in June 2013 provided an opportunity for an observational study evaluating the impact of this new intervention upon the social skills and behaviour of the Young Persons residing there. The significance of the TFH model for Youth Horizons will be known to the current readership and this study will be the first New Zealand based study of the TFH model. This paper represents an initial report for a Youth Horizons readership; as such some details will be omitted such as a description and literature review of the TFH methodology and the description of statistical methods will be slightly simplified.

Method

Participants and setting

Three young persons participated in this study, this being all of the young persons residing at the Te Whare Tumanako o Kia Puawai residence for a period spanning one month before and after the introduction of the Teaching Family Model. All had been referred by Child Youth and Family for residential placement with Youth Horizons on the basis of disruptive behaviour such as to render the child in need of care and protection. Parents were all briefed by the researcher and provided written consent for their children to participate.

“Tony” (not his real name) was 10 years of age and had previous diagnoses of Attention-Deficit Hyperactivity Disorder and Conduct Disorder. Tony had been excluded from school for problem behaviours and had not attended school in over a year. A rating scale measure of Tony’s behaviour prior to entering the residence placed him at the 95th percentile for age for aggressive

and rule breaking behaviours. His most significant challenging behaviour at the onset of the study was extreme behavioural escalations of up to an hour or more in duration involving property damage and aggression to staff. Over the month prior to the commencement of TFH Tony averaged 4-5 behavioural escalations a week.

"Tyson" was 9 years of age and had a previous diagnosis of Attention-Deficit Hyperactivity Disorder. Although Tyson had no other formal diagnoses, in a behavioural questionnaire completed three months prior to the commencement of this study Tyson's father indicated that Tyson exhibited four of eight behaviours associated with Oppositional Defiant Disorder (four are required for DSM V diagnosis) and three of fifteen behaviours associated with Conduct Disorder (three required for diagnosis). Another rating scale measure of Tyson's behaviour completed by his father two months prior to the commencement of the present study placed him at over the 98th percentile for age for aggressive and rule breaking behaviours. Like Tony Tyson's most significant presenting issue in the residence was behavioural escalations involving property damage and aggression to staff and other Young Persons. Over the month prior to the commencement of TFH Tyson averaged 3-4 behavioural escalations a week.

"Philip" was 15 years of age and had previous diagnoses of Aspergers Syndrome & Conduct Disorder. A rating scale measure of Philip's behaviour prior to entering the residence completed by his mother placed him at the 98th percentile for age for aggressive and rule breaking behaviours. Although not formally excluded Philip had been asked not to return to his school in 2013 because of his challenging behaviour the previous year. Over the month prior to the commencement of TFH Philip was involved in behavioural escalations from once a fortnight to once a week.

The setting was a five bedroomed group home in South Auckland: Te Whare Tumanako o Kia Puawai (TWT). This is a non-secure residence (no fences or locked doors) in appearance very similar to an ordinary suburban house. Staff were drawn from a pool of approximately 12 rostered staff who did eight hour shifts. The house was staffed 24/7. Staff were experienced Specialist Youth Workers; other staff attached to the residence included a House manager, Social worker and Psychologist.

Prior to the introduction of TFH the existing treatment model at Te Whare Tumanako o Kia Puawai was structured around a token economy points system and case-management process derived from a treatment fostercare model, together with a range of behavioural methods such as contingency contracts, redirection and planned ignoring. For convenience we will refer to this model as the "previous" model.

Design

In single-case research (SCR)* the most powerful experimental designs involve either a return to baseline after the introduction of the treatment (ABA or ABAB designs) or the sequential introduction of the treatment across participants or behaviours (multiple baseline designs). In the current context the former might involve the introduction of the TFH model for a period followed by a return to the previous model for a period before re-introducing TFH, and perhaps repeating this cycle. The latter might involve delivering the TFH model to one or more boys in the residence whilst simultaneously delivering the previous model to other boys in the household. Difficulties with these options include switching from one complex treatment model to another and then back again on set dates, the ethical sensitivity of removing an effective intervention and replacing it with a hypothetically less effective one and the technical challenge posed by the need to deliver different interventions consistently to different young persons who are in close proximity i.e. sitting at the same table. For these reasons and others the current study was restricted to a quasi-experimental "AB" design. In this design data is collected for a period of time prior to the introduction of an intervention (the baseline or "A" phase) and for a period of time after the introduction of the intervention (the treatment or "B" phase). Three simultaneous AB designs were planned, one per participant, whereby data was to be collected for approximately a month before and after the introduction of the TFH model. Data collection would then cease, although the TFH intervention would continue. Data collection would recommence for ten days, three months after the introduction of the TFH model *or* ten days prior to discharge, whichever came first, in order to verify the degree to which any behaviour change had been maintained over a longer term.

Measures

Single case research is typically based on data collected via direct observation by trained observers. Because of the cost and effort observations are typically conducted in sessions and are rarely conducted for more than 1-2 hours a day for any extended period. Three weaknesses of direct observation for Young Persons with conduct problems are firstly, young persons are likely to alter their behaviour in the presence of unfamiliar people, secondly many conduct problem behaviours are covert and thirdly many non-covert conduct problems occur at a very low rate – once or twice a week, or less. Even an observation schedule of 2 hours a day will only cover a little over 10% of the young persons waking hours and it is likely that the majority of low frequency behaviours will not be observed. For this reason an alternative method was required. A measure based on Chamberlain and Reid's Parent Daily Report (PDR)¹ was developed and named the "Daily Behaviour Checklist" (DBC). Twenty five items based on items from published rating scales for social skills

* In Psychology and Education, research where data is collected and analysed on a participant by participant basis and not aggregated.

and disruptive behaviour as well as DSM IV criteria for Oppositional Defiant Disorder and Conduct Disorder were used (see Appendix 2 below). Of these, fifteen items reflected negative or anti-social behaviours such as "angry", "argued with adults" and "bullying/intimidation" and ten represented positive or prosocial behaviours such as "made a compromise during a conflict" and "asked permission". An additional five prosocial behaviours directly taught as part of the TFH intervention were also included such as "accepted no" and "raised a concern". Twice a day, mid-afternoon towards the end of the day shift (7am to 3pm) and after the young persons had gone to bed, a residence staff-person would rate each of the thirty behaviours as either "occurred" or "not-occurred" over the previous eight hour period. Behaviours occurring overnight were rare and were not assessed. The staff completing the form were able to use all sources of information such as the report of other staff or written documentation such as points cards or school cards. The total positive and negative behaviours were summed to yield "positive" and "negative" scores between zero and fifteen. In contrast to direct observation the DBC allows data collection for every waking hour for the child for extended periods, observation by regular house staff who will engender minimal observer reactivity and the recording of covert behaviours disclosed by evidence but not observed, e.g. the theft was not seen but the object was found hidden in the boys room. The disadvantage of this "indirect observation" method is that it is likely to be less reliable than direct observation. Incident data, which consists of descriptive reports of moderate to major behavioural incidents written by the staff who witnessed the event, was also collected. Incident data is routinely collected in all Youth Horizons' residences.

Inter-rater reliability

The observers in this study were the regular residence staff. The DBC measure was explained in person by the researcher to the staff. A minority of casual staff were unavailable for these briefings owing to their irregular hours; these staff were advised by senior staff who had attended the briefings. The DBC was considered to be akin to a rating scale such as the Connors or Child Behaviour Checklist where observer training is not required. Nonetheless guidelines for staff were printed on the back of each DBC form.

On 40 or approximately 10% of the 414 measurement occasions the researcher telephoned the residence and required an arbitrarily selected staff person to complete a DBC over the phone for a young person who had previously been rated for that period by another staff person. Staff were specifically instructed not to look at the previous questionnaire and the two different informants' DBC's for the same Young Person over the same period were subsequently compared by the researcher. Reliability was calculated as item by item inter-observer agreement multiplied by 100 to yield a percentage. The average agreement across all observations was 78% with a range from 100% to 57%. Although modest by the standards of many psychometrics this value compares well with results obtained in the original

PDR standardisation study where mean inter-observer agreement between parents was equivalent to 53%[†]

Internal consistency

Cronbach's alpha was also calculated using every fifth repetition of the DBC in the study in order to minimise serial correlation between repetitions of the scale. Alpha is a measure of the extent to which the individual items of an assessment measure the same underlying factor as the scale as a whole. An overall alpha of 0.90 was obtained indicating a high degree of internal consistency. Negative behaviours were reverse coded so that the positive and negative total scores were in the same direction. Those items with the strongest correlations with the remainder of the scale related to anger, defiance, non-compliance and blaming others and were thus closely related to symptoms of Oppositional Defiant Disorder. However this data must be interpreted with caution as serial repetitions of the DBC on the same individuals were used and this may have inflated inter-item correlations.

Analysis strategy

The decision was made prior to the commencement of the study to augment the visual analysis of graphed data, which is the standard approach in Applied Behaviour Analysis and in single-case research in general, with statistical measures and calculation of effect sizes. The arguments for and against statistical measures in single-case research will not be repeated here however the following issues were considerations.

As the baseline period was itself a behavioural intervention of some probable efficacy it was anticipated that the incremental effect of TFH could be small and may benefit from a numerical measure of effect size and statistical tests of significance. In other words it was anticipated that it may prove to be difficult to detect any change due to the TFH intervention based on a visual analysis alone.

The available design, a quasi-experimental repeated AB, has poor internal validity and may not allow a convincing depiction of experimental change, or the lack of it; statistical methods may serve to clarify the degree and nature of this change.

The use of statistical measures in single-case research is rare and their value or usefulness in the context of an applied study will be of some interest to researchers.

[†] Chamberlain and Reid (1987); inter-parent agreement was cited as a correlation, $r = 0.02$, this is equivalent to 53% agreement in a 30 item scale.

A difficulty with the use of statistical methods in single-case research is that there is no single agreed method and different techniques have different advantages and strengths. For this reason three statistical methods were used as described below. The intention was to use all sources of information, statistical and visual, to formulate an overall description of the degree and significance of any change due to the TFH intervention.

Statistical methods

- 1) Standardised mean difference, Cohen's d .² This is an effect size index which is widely used in multiple group research and meta-analysis and has been recommended as an effect statistic for use in SCR³. Conventionally values of d of 0.2 indicate a small effect, 0.5 a medium effect and 0.8 a large effect however obtained values of d are likely to be significantly larger where the data are repeated measures as discussed below. The formula for d is

$$d = \frac{M_t - M_b}{SD_{pooled}}$$

Where M_b is the mean of the baseline phase, M_t is the mean of the treatment phase and SD_{pooled} is the pooled standard deviation of the two phases. Cohen's d was calculated for the comparison between baseline and the treatment phase and for the comparison between baseline and the three month follow-up (or just before discharge) period.

- 2) The regression based method due to Huitema et al.⁴. A number of techniques based on linear regression have been proposed for the analysis of single case research^{5,6} and have been used in meta-analyses⁷. Huitema's method is more recent and more straightforward to calculate than others. The advantage of regression approaches is that unlike standardised mean differences the method can evaluate changes in both level and trend and can control for baseline trend, a common confound. This method will indicate whether changes in level and trend are statistically significant and it will also yield an alternative estimate of d . A weakness of this approach is that it makes statistical assumptions which may not be realised in time series data hence the values obtained may be biased. Because this method requires successive data points – no significant gaps in the data series – regression statistics were calculated for the baseline-treatment contrast only (not for follow-up).
- 3) Parker et. al.'s Tau-U statistic⁸. Tau-U is a newly developed statistic based on Kendall's Tau and the Mann-Whitney U test. Unlike

regression methods Tau-U is a non-parametric statistic which is theoretically valid for time series data. Essentially Tau-U is an index of overlap in a matrix where every baseline data point is compared with every treatment data point. For instance if both are four observations in duration there will be a 4 x 4 matrix with 16 comparisons. Tau-U is the ratio of the number of pairwise comparisons where the treatment measure is larger (or smaller) than baseline over the total number of pairs. This value has a known distribution and tests of significance can be derived. Values of Tau-U approaching 1.00 indicate minimal overlap and a marked difference between phases, values approaching zero indicate a high degree of overlap and minimal difference between phases. The method as described by Parker et al. also allows for adjustment to incorporate the effect of baseline trend. It was considered questionable whether calculating Tau U for the contrast between baseline and follow-up – which involved a gap of several weeks of no data – was valid and so Tau was calculated for the baseline-treatment contrast only.

Cohen's d was calculated using Excel and Huitema's regression using STATA version 13; Bishop's statistic was calculated using a combination of the two and to ensure accuracy all Tau calculations were repeated using an online calculator for Tau-U⁹.

Results

Tony Negative behaviour

Tony's negative behaviour score during baseline was variable with changes of up to 11 points on a 15 point scale between successive measurements (Fig. 1 below). There was no visually apparent trend during baseline and a statistical test for trend was not significant (Table 1 "Tau BL trend"). Tony achieved a marked decrease in negative behaviour score from an average of 7.4 in the baseline period to 4.47 during the intervention period, a change equivalent to a Cohen's d of 0.84, ostensibly a large effect and clearly apparent in the graphed data (see Fig 1). The Huitema regression concurred with a similar effect size of 0.88 and the test for a significant difference between the mean baseline and treatment scores (change in level) was significant ($t(1,118) = 2.26, p=0.03$). A test of the difference in trend between baseline and treatment was not significant ($p=0.17$). Bishops method yielded a Tau of 0.45 and this outcome was statistically significant ($p<0.01$). Tony's gains during the treatment period were maintained and improved at the three month follow-up with a further decrease to a score of 3.00 equivalent to a Cohen's d of 1.27 when contrasted with baseline. Overall Tony's negative behaviours decreased following the commencement of the TFH methodology, a change that was large, assessed as statistically significant by two of two methods and maintained at follow-up.

Tony Positive behaviour

Tony's positive behaviour score during baseline was variable with changes of up to 12 points on a 15 point scale between successive measurements. There was no visually apparent trend during baseline and this and a statistical test for trend was not significant. Tony's positive behaviour score increased from a mean of 5.58 during baseline to 7.86 during the treatment phase a change equivalent to a Cohen's d of 0.65, a medium effect size.

Figure 1: Tony Negative behaviours

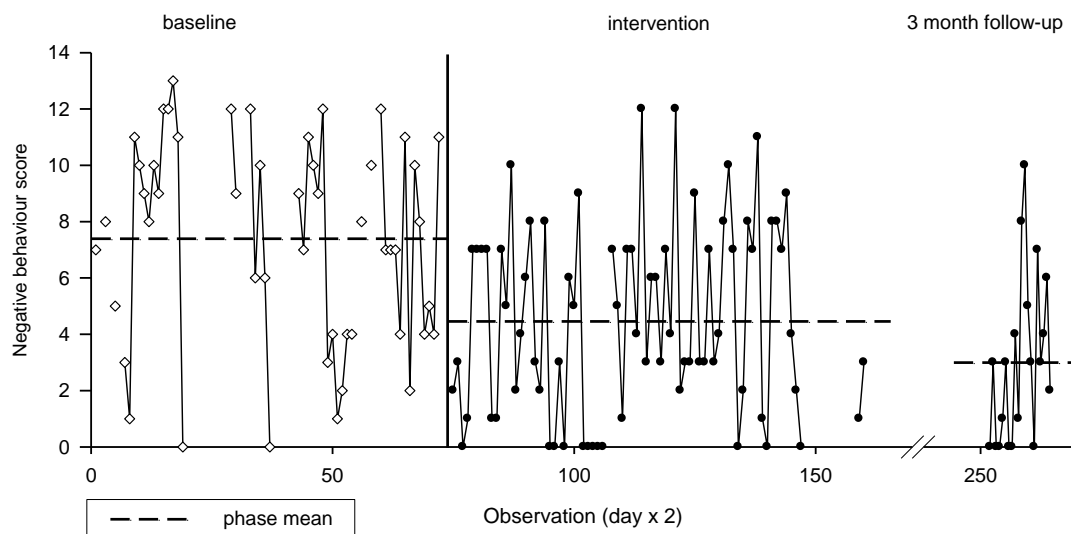
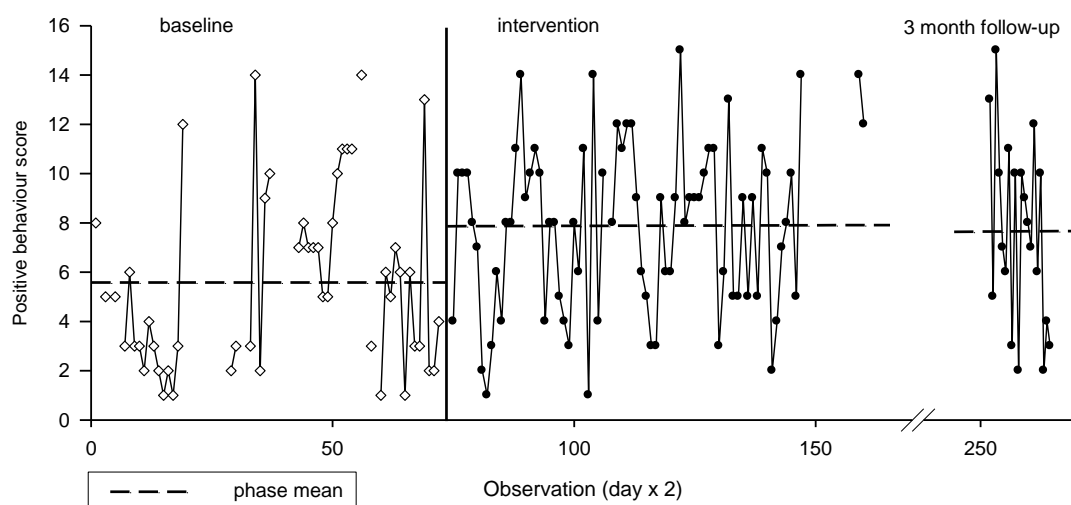


Figure 2: Tony Positive behaviours



The Huitema regression method yielded a similar value of d , 0.74 but did not indicate a significant difference in either level or slope between baseline and treatment. In contrast Tau U for this comparison was 0.37 which was significant ($p < 0.01$). Tony's mean positive behaviour score at follow-up was 7.65, similar but slightly lower than the mean treatment score and Cohen's d for the baseline-follow-up comparison was 0.56. Overall Tony's positive behaviour score increased to an extent equivalent to a medium effect size and this was assessed as statistically significant by one of two methods. Treatment gains were largely maintained at follow-up.

Figure 3: Tyson Negative behaviours

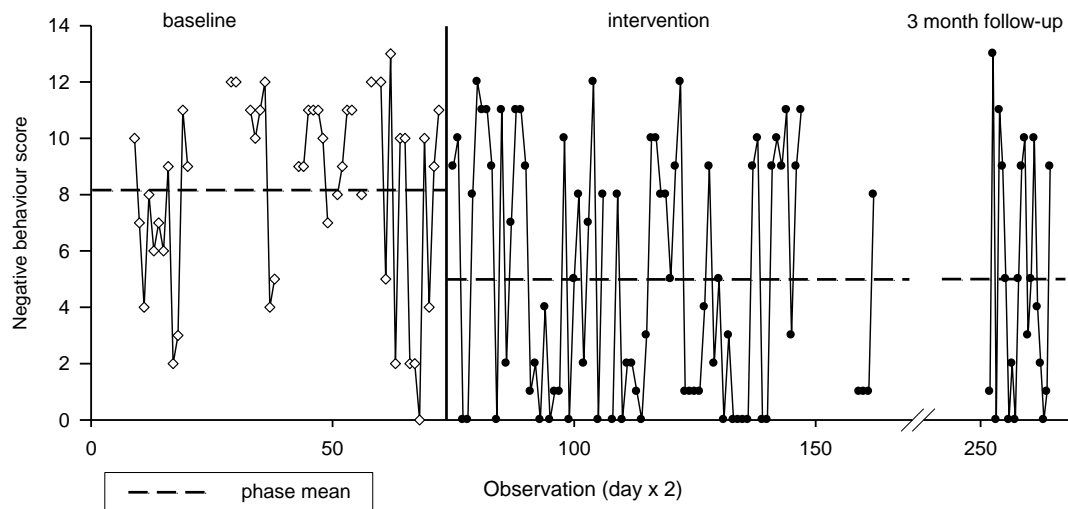
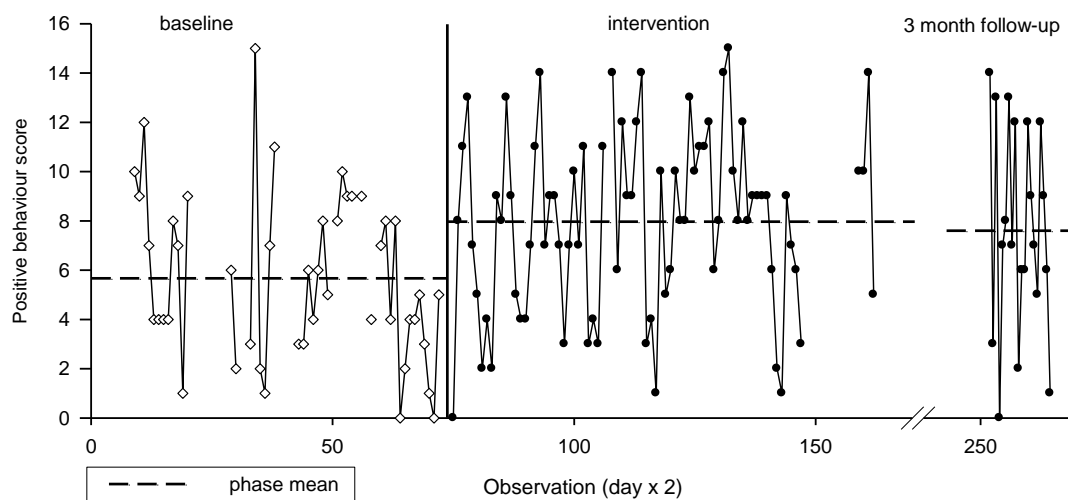


Figure 4: Tyson Positive behaviours



Tyson Negative behaviour

Tyson's negative behaviour scores during baseline were variable with no discernible or statistically identified trend. A baseline mean of 8.17 decreased to 4.99 during the intervention phase a change equivalent to a Cohen's *d* of 0.79, a large and visually apparent effect. The regression method was broadly consistent yielding a *d* of 0.76 but no significant change in level or trend from baseline to treatment. Tau-U for this contrast was significant (Tau 0.43, $p < 0.01$). Gains were maintained and slightly improved at follow-up with a mean score of 4.95 equivalent to a Cohen's *d* of 0.87 relative to baseline. Overall Tyson's negative behaviour score decreased to an extent equivalent to a large effect size and this was assessed as statistically significant by one of two methods. Treatment gains were maintained at follow-up.

Tyson Positive behaviour

Tyson's positive behaviour during baseline was also variable and evidenced a mild non-significant trend towards lower scores over time ($p = 0.09$, Table 1). This finding was supported by the near-significant change in slope from baseline to treatment as indicated by the Huitema method ($t_{1,65} p < 0.06$). These effects are somewhat visible in Figure 4 as gradually decreasing scores from observations 1 to 72 with a hint of improving scores over the first three quarters of the treatment phase. Although not statistically significant the presence of this trend change is consistent with an effect due to the TFH method as a deteriorating trend during baseline is reversed and becomes a trend towards improvement during the intervention phase. Tyson's mean positive behaviour score increased from 5.67 to 7.96 from baseline to treatment yielding a *d* of 0.65, a medium to large effect size. Cohen's *d* from the regression method was very similar, 0.71. This change was statistically significant according to both the Huitema method ($t_{(1,118)} 2.21, p < 0.03$) and Tau U (Tau 0.41. $p < 0.01$ – corrected for baseline trend). At three month follow-up Tyson's mean score was 7.60, equivalent to a *d* of 0.53, a slight deterioration relative to the treatment phase. Overall Tyson's positive behaviour score increased to an extent equivalent to a medium to large effect and this was assessed as statistically significant by two of two methods. Treatment gains were maintained at follow-up.

Philip Negative behaviours

Philip's negative behaviour scores during baseline were also variable with changes of up to 12 points on successive measurement occasions. There was no visually apparent trend during baseline and no significant trend as assessed statistically. Philip's mean baseline score of 2.60 decreased to a mean of 1.77 during the treatment period equivalent to a Cohen's *d* of 0.28, a small effect. There were some indications of a trend towards deterioration over the course of the treatment period as evidenced by a series of gradually higher peak scores (see Figure 5).

Figure 5: Philip Negative behaviours

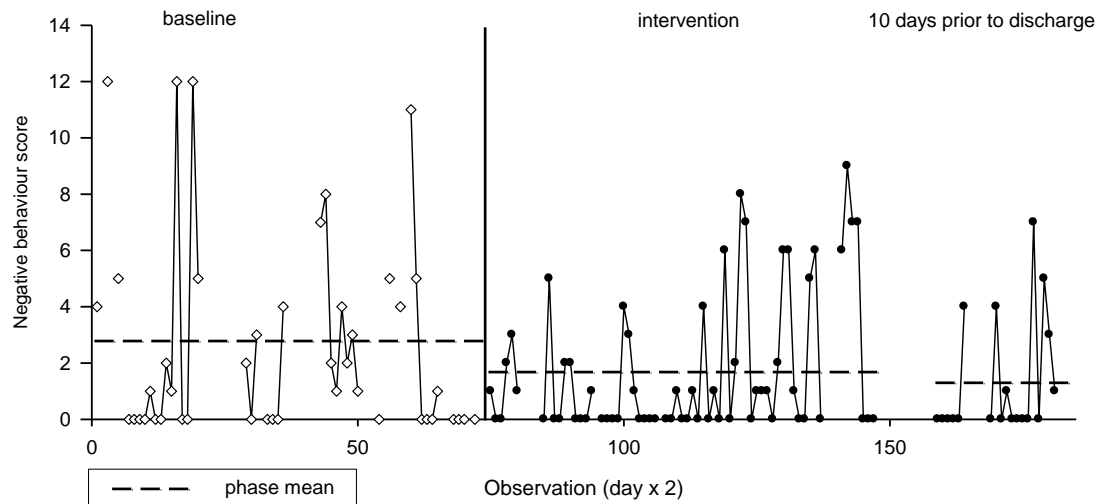
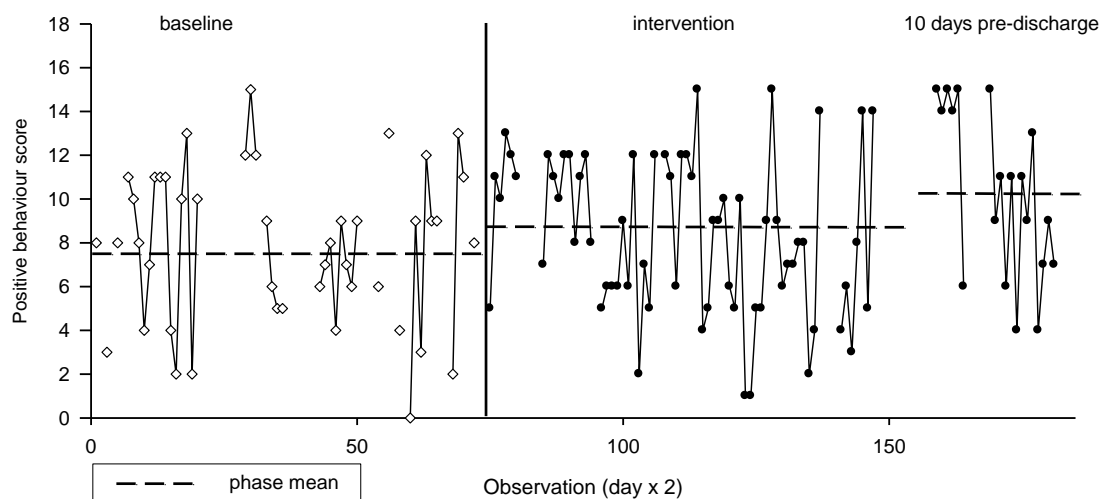


Figure 6: Philip Positive behaviours



In terms of the Huitema regression method the change in slope was significant ($t(1,118) 2.42, p < 0.02$) but the change in level was not. It is clear from the graph however that the change in slope is from close to nil in baseline to deterioration in the treatment, a change in the opposite direction from that desired. The regression method indicated a d of 0.47, a medium effect size, however this does not represent a treatment effect as some of the change is in the "wrong direction". The value of Tau U obtained was 0.10 indicating a high degree of overlap between baseline and treatment and this was not statistically significant ($p = 0.38$). As Philip was due to be discharged data was collected for ten days commencing shortly after the completion of data collection for the treatment phase. Philip obtained an average negative

behaviour score of 1.31 representing a Cohen's d of 0.40, a slight improvement relative to the treatment period and a small to medium effect relative to baseline. Overall Philip's negative behaviour score decreased slightly to an extent equivalent to a small effect and this was assessed as not statistically significant by two of two methods, with the exception of a significant change in slope in the direction of deterioration. Treatment gains were slightly improved upon during the period just before discharge to a degree equivalent to a medium effect relative to baseline.

Philip Positive behaviours

Philip's positive behaviours during baseline were variable and displayed no visually apparent or statistically ascertained trend. His mean positive behaviour score improved slightly from a mean of 7.67 during baseline to 8.33 during the intervention period; this translated to a Cohen's d of 0.19, a small effect. In terms of the Huitema regression method there were no significant changes in level or slope although the change in level was near-significant ($p < 0.08$) and the overall value of d obtained was 0.43, a medium effect. The contrast between baseline and intervention as assessed using Tau U was not significant (Tau 0.10, $p = 0.39$). Philip's positive behaviour score during the pre-discharge period was improved compared to the intervention period with a mean of 10.26 equivalent to a d of 0.74 relative to baseline, a large effect, and this change was somewhat visually apparent in the graph. Overall Philip's positive behaviour score increased to an extent equivalent to a small or medium effect and this was assessed as not statistically significant by two of two methods. Treatment gains were improved upon during the period just before discharge to a degree equivalent to a large effect relative to baseline.

Overall test of statistical significance

The Tau-U technique allows for the aggregation of individual tests across participants so that an overall or omnibus level of statistical significance can be determined. This provides an indication of whether the outcomes for all three participants considered together are statistically significant. In the case of the negative behaviour scores the omnibus value of Tau-U was 0.33, which was statistically significant ($p < 0.01$). With regards to the positive behaviour scores the obtained value of Tau-U was 0.28, also statistically significant ($p < 0.01$). Note that the values of Tau-U obtained were intermediate between the values obtained for Tony and Tyson on one hand and Philip on the other. Although as we have seen Philip's changes in negative and positive behaviour yielded small values of Tau-U which were not significant this was not sufficient to render the outcomes non-significant for the trio considered as a group.

Analysis of directly and indirectly taught behaviours.

Prima facie the only behaviours directly taught or consequenced in this study were a) TFH skills such as *accepting no for an answer* and *asking permission*. and b) negative behaviours consequenced via response cost (points fines) – which is also an explicit part of model for managing behavioural escalations. Thus the presumed mechanism for the acquisition of non-TFH positive skills such as *considerate of others* and *tolerated a peer when he was annoying* – and indeed the countless number of social skills not listed in the TFH manual – must then be response generalisation; the spontaneous performance of responses which are functionally equivalent to the trained response. For instance a child taught to *accept criticism calmly* may, without any explicit training, begin to *respond calmly when teased or provoked*. On this basis certain predictions can be made as to the relationship between the three classes of behaviour monitored via the DBC measure; TFH skills, other non-TFH positive behaviours and negative behaviours. Firstly, the TFH skills and other positive behaviours would be expected to be positively correlated to a moderate or strong degree; this was the case, positive behaviours and TFH skills were correlated to a moderate and similar degree within participant (r 0.61 to 0.63). Both would be expected to be negatively correlated with the negative behaviour score within participant, and this was the case (r -0.23 to -0.54). Given that in a sense change in TFH skills is hypothesised to drive change in the non-TFH positive skills effect sizes for the former would be expected to be at least equal and possibly larger than for the latter. This was also the case; for the baseline treatment comparison, Cohen's d , TFH vs. non-TFH respectively; Philip 0.22, 0.14; Tony 0.60, 0.44 and Tyson 0.69, 0.59. Thus in all cases improvements in the directly taught TFH skills were somewhat larger than improvements in those positive skills which were not directly taught, consistent with the hypothesis that the non-TFH skills are acquired via response generalisation. This outcome is not conclusive however as there are other potential explanations; for instance staff might have been interpreting the TFH skill descriptions loosely and reinforcing a variety of positive behaviours. However three eventualities which would falsify this hypothesis did not occur; minimal change in TFH skills alongside marked change in non-TFH positive behaviour, minimal change in non-TFH positive behaviour alongside marked change in TFH skills behaviour, or nil or minimal change in both.

Summary and discussion

An AB quasi-experimental design repeated simultaneously across three participants was used to contrast the effect of the Teaching Family Homes behavioural model (intervention phase) with a generic behavioural model (baseline phase). Two of the participants obtained visually apparent improvements in both positive and negative behaviours (increases and decreases respectively) equivalent to moderate to large effect sizes and two of the four baseline-treatment contrasts for these two individuals were assessed as statistically significant by a linear regression method and four of

four were assessed as statistically significant by a non-parametric statistical method. Improvements were maintained at a three month follow-up. A third participant achieved improvements in positive and negative behaviours which were somewhat apparent in the graphed data and equivalent to medium to small effect sizes which were assessed as not significant by both statistical methods. These improvements were maintained or increased at an early follow-up which commenced a few days after the intervention data-collection period as the young person was due to be discharged.

All significant behavioural escalations in the residence were recorded by staff via written incident reports. This process was standard practice and was conducted independently from the DBC data collection. The baseline was 36 days in duration thus incident totals for the 36 days before and after the introduction of the TFH system were tallied as follows; Tony, 18 incidents before, 2 incidents after; Tyson, 15 incidents before, 12 incidents after and Philip, 5 incidents before, 2 incidents after. As incidents were reduced in all cases this data provides independent support of the validity of the DBC data and additional support to the contention that the TFH model lead to a reduction in negative behaviours relative to the baseline condition.

Thus in two out of three cases the introduction of the TFH model coincided with decreases in the incidence of problem behaviours as well as increases in the rate of positive pro-social behaviours relative to a pre-existing behavioural programme. In the case of the third young person there was a significant deterioration in trend for negative behaviours, however the overall change in the mean level of behaviours represented small to large (depending on the measure and period considered) improvements in behaviours relative to baseline and thus are not inconsistent with the outcomes for the first two participants. The achieved changes in this study were particularly notable as the baseline condition was itself an intervention of some probable efficacy as it included evidence based components such as a token-economy. The maintenance of treatment effects at follow-up is not maintenance in the technical sense of post-treatment effects as the TFH intervention was still in place however this finding provides additional evidence that the changes measured over the initial intervention period were not due to chance as similar improvements were detected during a second observation period removed in time from the first. Further, the follow-up data provides evidence that the effect of TFH was not due to a transient "novelty" effect which waned over time.

Notwithstanding the results as presented there are a number of weaknesses in this study.

Implementation problems

There were a significant degree of missing observations especially during the baseline phase. This was almost certainly the result of using regular staff in an applied setting where each had very many daily duties and tasks in

addition to the completion of the DBC observation sheets, a task for which they were given no additional remuneration. Most staff had multiple recording and reporting tasks and it clearly took some time from the initiation of the study for the collection of the required six DBC's a day to become routine. There was very little data loss upon the commencement of the treatment phase. Not every missed observation was an omission; Philip was having regular weekend home-leave by the end of the study. An advantage of having a comparatively long baseline period is that even with missed observations the levels of behaviour were well sampled with 45 to 50 observations per Young Person over the baseline phase.

The Daily Behaviour Checklist is not a well-established measure and its psychometric properties – apart from in the present study – are unknown. The mean inter-observer reliability of 78% obtained was just below the generally accepted minimum standard of 80% and some observations on some occasions were quite unreliable (inter-observer agreements were as low as 57% on occasion). Also the percentage of observation occasions assessed for reliability was also below the generally accepted minimum of 20% of occasions. However as noted above the inter-observer reliability was no worse than inter-parent agreement quoted for the Parent Daily Report in an early study and this measure has been used in a number of published studies.

Design limitations

In terms of expected methodological limitations the AB design used is not robust against threats to internal validity and cannot rule out certain alternative explanations for apparently favourable results. It is possible that an effect or influence other than TFH produced the outcomes achieved (this is known as "history"). For instance a significant staffing change (extra staff, more skilled staff), departure of a particularly disruptive young person or the coincidental introduction of some other intervention (medication, Cognitive Behaviour Therapy) might have been responsible for the changes in outcome. If it is accepted that levels of problem behaviour during baseline were exceptionally high then any subsequent decrease may be due in part to the purely statistical tendency for extreme levels on any variable to become less extreme over time (regression to the mean). The counter-argument is that there is evidence of some quite rapid change in some of the graphs; the level change in Tony's negative behaviour score and the rate of zero scores for Tyson's negative behaviour (one during baseline, 17 during the intervention phase) and this did not coincide with any known change in procedure, staff or treatment beyond the introduction of TFH.

Another possibility is that the baseline period may have coincided with a period of increased problem behaviour in the residence for whatever reason. This has some currency as two boys departed the residence in the course of this study due to high levels of behavioural outbursts involving these young persons as well as the study participants. One of these boys "T" was only in the residence for seven days coinciding with data-points 14-28, however

there is no discernible pattern in the data for any of the participants such as a decrease in negative behaviour scores following T's exit. A second boy "B" had been in the residence for some months and was discharged on the day coinciding with data point 88 in the early part of the intervention phase; similarly there is no clear evidence of this having an impact on outcomes. For instance, Tony's negative behaviour score appears to have decreased prior to B's departure and if anything his score deteriorates over the remainder of the intervention phase. Philip's negative behaviour appears to increase following data point 88, which is the opposite of what would be expected on the assumption that the departure of B contributed to the apparent treatment effect. Further, Tyson achieves three nil scores (no problem behaviours) within a week after the introduction of TFH and before the departure of B whereas he had achieved only a single nil score day over the course of the baseline period. Despite these indications, because the design does not include a return to the baseline phase the possible influence of other residents in the house, or other unknown factors, cannot be ruled out.

A further limitation of the present design is the low numbers of participants involved, three; because the participant group is small and not a random sample from any larger group the generalisability of the results is limited. The partial solution to this is to where possible replicate the study across different settings and participants.

Statistical methods

In the present study the Huitema method yielded comparable values of Cohen's d to the standard formula in most cases and appeared consistent with the graphed outcomes. It recorded large effect sizes as statistically significant or near significant but not otherwise. The Tau U and Huitema statistics agreed in four of six cases as to statistical significance; the two exceptions (Tony positive and Tyson negative behaviours) were both judged significant via Tau- U and non-significant via Huitema's method. As the values of Cohen's d were medium to large the Tau- U outcomes were not unrealistic. This supports the validity of the Tau- U statistic and hence the overall finding of treatment effects due to the TFH intervention. What would *not* support this thesis would be indications that the Tau- U statistic yielded significant outcomes for small values of Cohen's d (Type 1 error, false positives) but this was not found.

Finally the descriptions of the values of Cohen's d (0.2 is "small", 0.5 is "medium", 0.8 is "large") are based on the accepted interpretation of d for group comparison research, that is, for comparisons between an intervention group and a control group. There is however evidence from the meta-analysis literature that effect sizes calculated pre-post, as in the present study, may produce values of d 20% to 60% larger than comparable between groups studies for the same intervention^{9,10}. Although the values of d obtained in a pre-post study will not change, the meaning of a given value of d may vary. In the present case the average value of d for the baseline treatment

comparison across all participants, negative and positive behaviours was 0.57, a medium effect. If we assume that this involves a 20% to 60% “inflation” relative to group designs a corrected value of d would range from 0.47 to 0.35, a medium to small effect. Having said that, it can be argued that the “inflation” findings are based on single-group pre-test post-test designs with one measurement point per participant and the failure of these designs to control for confounds such as maturation[‡], which is controlled for in a repeated measures design where there is a baseline which can be examined for any pre-existing trend. In the present study larger values of Cohen’s d generally corresponded with statistically significant differences as evaluated by Tau-U and visually apparent changes in the graphed outcomes hence the conventional evaluation of d whereby the mean value obtained corresponded to a “medium” effect ($d = 0.57$) is plausible.

Further research

The following relates to research opportunities within Youth Horizons Trust.

The question of response generalisation from directly taught TFH skills will require a traditional SCR design and direct observation in order to monitor the rates per minute or hour of taught and untaught skills or behaviours. This research may be conducted over a relatively short period of time (one or two weeks) and may be suitable for a student project.

Ongoing refinement of the Daily Behaviour Checklist should be undertaken. Although the Parent Daily Report is widely used in research and practice and is the mainstay of daily data collection in MTFC much of its content seems oriented towards mild or minor behaviour problems and it also has no positive items. For instance PDR items include “pantswetting”, “complaining” and “pouting” and there is no item for absconding.

Compared with other models such as MST and MTFC TFH is relatively under-researched and any reasonably competent study documenting outcomes in TFH facilities will be a contribution to the literature. Because the present study only involved three young persons further replications will be useful and significant, especially where a caregiver or residence transitions from a different model of care to TFH. Where a young person is placed with a different organisation or a parent prior to entering a TFH facility it may be possible to gather baseline data via DBC or PDR conducted over the phone every 1-3 days. Where amongst a cohort of YHT caregivers some offer TFH and others do not there is a clear opportunity for a comparative study albeit one with certain weaknesses, in particular the possibility that the TFH and non-TFH caregivers and Young Persons are systematically different prior to any intervention.

[‡] Any biological or psychological process which varies systematically independent of the intervention concerned. For instance, in interventions for children, a change in outcomes due to age.

As a large pilot study of TFH similar to the current FFT and MTFC research is a distant prospect at this time, and a control group study even further afield, the next level of achievable research apart from that mentioned above would probably be a non-concurrent multiple baseline across subjects design whereby for a series of new entrants to a particular residence they are initially placed on a generic non-TFH model such as the "previous model" in this study, for a period of time of 2, 4 or 6 weeks before moving to the TFH model. If three or more young persons enter a facility at the same time the study could be conducted concurrently. For a given Young Person who has transitioned to TFH his progress can be evaluated both relative to his own baseline and relative to those other Young Persons still in their baseline phase. There is an obvious sense in which a study like this would interfere with the normal functioning of TFH but it would be ethically relatively innocuous as all Young Persons would receive the preferred intervention. Such a study would require careful planning and probably ethics approval but if staff are the primary observers, minimal funding beyond the time of the researcher.

References

1. Chamberlain, P., & Reid, J. B. (1987). Parent observation and report of child symptoms. *Behavioral assessment* 9, 97-109
2. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (1st ed.). Hillsdale, N.J.: Erlbaum.
3. Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research *Single-case research design and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc; England.
4. Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement*, 60(1), 38-58.
5. Center, B. A., Skiba, R. J., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education*, 19(4), 1985-1986.
6. Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy*, 31(6), 621-631.
7. Allison, D. B., Faith, M. S., & Franklin, R. D. (1995). Antecedent exercise in the treatment of disruptive behavior: A meta-analytic review. *Clinical Psychology: Science and Practice*, 2(3), 279-304.

8. Parker, R. I. V., K.J. Davis, J.L. Sauber, S.B. (2011). Combining nonoverlap and trend for single case research: Tau-U. *Behavior Therapy*, 42(2), 284-299.
9. Vannest, K.J., Parker, R.I., & Gonen, O. (2011). Single Case Research: web based calculators for SCR analysis. (Version 1.0) [Web-based application]. College Station, TX: Texas A&M University. Retrieved 8th October 2013. Available from singlecaseresearch.org
10. Carlson, K. D., & Schmidt, F. L. (1999). Impact of experimental design on effect size: Findings from the research literature on training. *Journal of Applied Psychology*, 84(6), 851-862.
11. Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods*, 6(4), 413-429.

PRELIMINARY DRAFT 2014

Appendix 1, Table 1

<i>Participant/ behaviour</i>	<i>Base- line mean</i>	<i>Treat- ment. mean</i>	<i>Follow- up/pre disch. mean</i>	<i>Cohen's d treat- ment</i>	<i>Cohen's d follow- up</i>	<i>Huitema d</i>	<i>Huitema level t value</i>	<i>Huitema level p</i>	<i>Huitema slope t value</i>	<i>Huitema slope p</i>	<i>Tau BL trend</i>	<i>Tau U BL-Tx</i>	<i>Tau U p</i>
Tony negative behaviours	7.40	4.47	3.00	0.84	1.27	0.88	-2.26	0.03*	1.37	0.17 ns	no	0.45	<0.01**
Tony positive behaviours	5.58	7.86	7.65	0.65	0.56	0.74	-0.24	0.81 ns	-0.25	0.80ns	no	0.37	<0.01**
Tyson negative behaviours	8.17	4.99	4.95	0.79	0.87	0.76	-1.77	0.08 ns	-0.49	0.62 ns	no	0.43	<0.01**
Tyson positive behaviours	5.67	7.96	7.60	0.65	0.53	0.71	2.21	0.03*	1.89	0.06 ns	P<0.09	0.41	<0.01**
Philip negative behaviours	2.60	1.77	1.31	0.28	0.40	0.47	-1.43	0.16 ns	2.42	0.02*	no	0.10	0.38 ns
Philip Positive behaviours	7.67	8.33	10.26	0.19	0.74	0.43	1.78	0.08 ns	-1.18	0.24 ns	no	0.10	0.39 ns

*p<0.05 **p<0.01 ns = not significant

Appendix 2**Daily Behaviour Checklist**

Young Person _____ Person completing form _____

Date _____

To be completed at the end of the shift

To the best of your knowledge to what extent did the young person do the following behaviours today ?

1 = occurred

0 = did not occur

	Am shift	Pm Shift
abscond/truant (30 min plus)		
accepted criticism calmly		
accepted no		
angry		
argued with adults		
asked for help		
asked permission		
blamed others for mistakes/misbehaviour		
broke household rules		
bullying/intimidation		
considerate of others		
damaged property		
Defiant		
got on well with other young persons		
greeted someone politely		
Helpful		
lost temper		
lying or cheating		
made a compromise during a conflict		
non-compliance with adult requests		
physical fight/ aggression		
raised a concern		
resolved a disagreement calmly		
responded calmly when teased or provoked		
stubborn sullen or irritable		
sudden change in mood		
Theft		
tolerated a peer when he was annoying		
took responsibility for his mistakes/misbehaviour		
volunteered for extra tasks		