A chance for change : building an outcome monitoring feedback system for outpatient mental health care

Jong, K. de

Chapter 3

A priori power analysis in longitudinal three-level multilevel models: an example with therapist effects

De Jong, K., Moerbeek, M. & Van der Leeden (2010).

Psychotherapy Research, 20(3), 273-284.

Abstract

Multilevel analysis (or hierarchical linear modeling) has become increasingly popular for the analysis of longitudinal data in psychotherapy research. Over the last few years, three-level longitudinal models have become more common in psychotherapy research, particularly in therapist-effect or group-effect studies. Thus far, limited attention has been paid to power analysis in these models. This article demonstrates the effects of intraclass correlation, level of randomization, sample size, covariates and drop-out on power, using data from a routine outcome monitoring study. Results indicate that randomization at the patient level is the most efficient, and that increasing the number of measurements does not increase power much. Adding a covariate or having a 25% drop-out rate had limited effects on study power in our data. In addition, the results demonstrate that sufficient power can be reached with small sample sizes, but that larger sample sizes are needed to prevent estimation bias for the model parameters and standard errors.

Introduction

Multilevel analysis (Goldstein, 2003; Hox, 2002; Raudenbusch & Bryk, 2002; Snijders & Bosker, 1999) has gained significant popularity as a statistical technique for the analysis of longitudinal data in psychotherapy research over the last decade. It has been used to analyze growth models of phenomenon such as patient progress and expected treatment response (e.g. Finch, Lambert, & Schaalje, 2001; Haas, Hill, Lambert, & Morrell, 2002; Lueger et al., 2001; Lutz, 2002; Lutz, Martinovich, & Howard, 1999; Lutz, Rafaeli, Howard, & Martinovich, 2002; Slee, Garnefski, Van der Leeden, Arensman, & Spinhoven, 2008), the dose-response relationship (Lutz, Lowry, Kopta, Einstein, & Howard, 2001) and group therapy (e.g. Haringsma, Engels, Van der Leeden, & Spinhoven, 2006; Tasca, Balfour, Ritchie, & Bissada, 2007). Multilevel analysis describes the class of methods employing hierarchical regression models, and is also referred to as hierarchical linear modeling, linear mixed modeling, random effects regression modeling and random coefficient modeling. These models explicitly take into account the hierarchical structure of the data (the fact that repeated measurements are nested within patients). The popularity of multilevel analysis in analyzing longitudinal psychotherapy data lies in its flexibility to handle missing data and unbalanced designs, its capacity to model individual growth trajectories (Van der Leeden, 1998), the within-subject covariance structure (Hedeker & Gibbons, 2006), as well as models with three or more levels.

In higher level models, an additional level at which the subjects are clustered (e.g. 'therapist' or 'organization'), is added to account for between-therapist or between-site variance. Three level models are becoming more common in psychotherapy research, especially in studies on group therapy and therapist effects (e.g. Elkin, Falconnier, Martinovich, & Mahoney, 2006; Lutz, Leon, Martinovich, Lyons, & Stiles, 2007; Okiishi, Lambert, Nielsen, & Ogles, 2003).

For a special section on therapist effects in *Psychotherapy Research,* data from the National Institute of Mental Health Treatment of Depression Collaborative Research Program (TRDCP, Elkin et al., 1989) were re-analyzed using two multilevel approaches. Employing a two-level (non-longitudinal) model with functioning at termination as the dependent variable, Kim, Wampold & Bolt (2006) showed significant differences between therapists in treatment results. In contrast, a three-level longitudinal model by Elkin, Falconnier, Martinovich & Mahoney (2006) did not find significant therapist variance. These contrasting results initiated an exploration of whether three-level longitudinal models are appropriate for psychotherapy research. Wampold & Bolt (2006) posit that, in psychotherapy research, the level of functioning at treatment termination is important; not the change pattern, arguing a move away from longitudinal models. However, there are serious drawbacks to both the completers sample analysis they suggest and the missing data imputation methods used in intent-to-treat analysis that

longitudinal analysis does not have (Crits-Christoph & Gallop, 2006).

One of the problems in the re-analyses of the TRDCP data was that the sample size was smaller than is generally recommended for multilevel analysis (Kim, Wampold, & Bolt, 2006; Soldz, 2006). Given that therapist-effects have been found in analyses with larger datasets (Lutz, Leon, Martinovich, Lyons and Stiles, 2007) it is possible that power might have been an issue in the TRDCP analyses. Therapist effects have usually been small in randomized controlled trials and small to medium in naturalistic studies (Crits-Christoph & Gallop, 2006). It seems that therapist effects are harder to detect in randomized controlled trials using treatment manuals (Crits-Christoph et al., 1991) and that therapist variance declines when therapists are trained (David M. Clark, personal correspondence). Yet, even a small amount of variance at the therapist level, can have a significant influence. It has been shown that ignoring a level of nesting in the data can have considerable effects on the estimated variances and power to detect treatment or covariate effects (Moerbeek, 2004) and can seriously inflate the Type I error rate and size of the treatment effect (Wampold & Serlin, 2000). Numerous researchers have shown that ignoring clustering can also lead to serious errors in interpreting the results of statistical significance tests (e.g. Nich & Carroll, 1997). As a result, in many cases, adding a third level that models between-therapist variance is necessary, regardless of whether therapist effects are the main topic of interest in the study.

Having nested data can strongly influence the power to detect a treatment effect. Unfortunately, many researchers still use power analyses that ignore the effect of nesting, even if they plan to do a multilevel analysis. The reason for this is that in multilevel analysis (a priori) estimation of power is complex, as it depends on a number of variables, including study duration, number of measurements, number of patients, number of therapists, the level of randomization and the intraclass correlation between the levels. Determining the sample size that is needed for sufficient power to detect a treatment effect is more complicated than in other approaches, because there are different sample sizes at the different levels: the number of measurements per patient (level 1), the number of patients per therapist (level 2) and the number of therapists (level 3). Moreover, it is necessary to use plausible values of the variance components in the model to get a proper estimation of power, and these are usually unknown.

Although some literature (e.g. Raudenbusch, 1997; Snijders & Bosker, 1993) and software solutions (Bosker, Snijders, & Guldemond, 1996) are available to estimate the power for two-level models, there is little information about power analysis for three-level longitudinal models. The Optimal Design software developed by Raudenbush and colleagues (Spybrook, Raudenbush, Liu, Congdon, & Martínez, 2008) is the only program we are aware of with an option to compute power for three-level models,

and this functionality exists only for limited designs. In this paper we will demonstrate a method to compute power for three-level longitudinal models. With limited financial resources it is very valuable to know whether it is more efficient to collect more measurements per patient, a larger number of patients with fewer measurements or a larger number of therapists, with a smaller number of patients per therapist. First, we will discuss factors that influence power in more detail. Then, the influence of these factors on power will be illustrated using naturalistic data from an outpatient setting.

Factors influencing power

Intraclass-correlation

Ignoring a significant intraclass correlation (ICC) by using traditional regression models can result in substantial estimation bias for parameters and standard errors (Goldstein, 2003). The ICC is defined as the degree of resemblance between micro-units belonging to the same macro-level unit and can also be interpreted as the fraction of total variability that is due to nesting (Snijders & Bosker, 1999). In multilevel modeling, although nesting is taken into account by the model, the size of the ICC strongly influences the power to detect treatment and covariate effects. For three-level models intraclass correlations are calculated for levels two and three. There are multiple ways to compute the level-two intraclass correlations (see Hox, 2002; Snijders & Bosker, 1999).

The effect of clustering can be clearly demonstrated by the design effect. The design effect is the ratio of patients required for modeling nested data, relative to non-nested data and is defined as $1+(k-1)ICC$, with k being the number of patients per therapist, and the ICC defined as the proportion of the total variance accounted for by the therapist level (Donner & Klar, 2000). For example, with an ICC of 0.1 and four patients per therapist the design effect would equal 1.3, meaning that approximately 30% more patients are needed for sufficient power than in non-nested data. The higher the ICC, the larger the design effect and the lower is the power to detect a treatment effect. In this example a higher ICC means that there is a larger difference between therapists, and that the therapist variance explains part of the variance of the treatment effect, thus reducing the power to detect that effect.

One of the major challenges in a priori power analysis for multilevel models is the need for plausible values of the variance components, and these are usually unknown.. Using ICC values from the literature will provide an idea of the extent to which the design-effect will decrease power. As the ICC depends on the outcome measure, it is necessary to base the ICC estimates on studies using the same outcome measures. In educational and medical research, there is an increasing amount of literature

reporting intraclass correlations for different outcome measures and covariates (e.g. Adams et al., 2004; Campbell, Mollison, & Grimshaw, 2001; Hedges & Hedberg, 2007). In psychotherapy research, such papers are desperately needed in order to perform sensible a priori power analyses.

Level of randomization

Three-level longitudinal models offer a choice of which level should be used to randomize to treatment conditions. In terms of statistical precision and power it is usually best to randomize at the lowest level possible (Moerbeek, 2005). In the case of three-level longitudinal models in psychotherapy research this is the patient level (level two), since the first level consists of the repeated measurements. A common design in randomized controlled trials in psychotherapy research is that patients are randomly assigned to treatment conditions, and therapists are recruited to either provide the experimental or control treatment. Although patients are randomly assigned to treatment conditions, statistically the randomization is at the therapist level, since within a therapist all patients receive the same treatment. The reason that this design is so common, is that it prevents contamination of the treatment effect. Contamination occurs when members of the experimental treatment condition influence members of the control condition. If therapists provide both the control (for instance treatment as usual) and the experimental treatment, they might unconsciously use techniques from the experimental treatment for their patients in the control group, thus contaminating the treatment effect. In case of contamination, randomization at the therapist level, actually might lead to higher power to detect an effect (Moerbeek, 2005). Alternative designs like pseudo-cluster randomization (Borm, Melis, Teerenstra, & Peer, 2005) and the split-plot design (Reise & Duan, 2003) have been developed to help to handle contamination, but have not been widely used so far.

Sample size

In three-level longitudinal models there are different sample sizes at different levels: the number of measurements per patient (*m*), the number of patients per therapist (*k*) and the number of therapists (*J*). Several combinations of *m, k* and *J* can result in an identical power to detect an effect, but a minimum sample size at each level is necessary for accurate estimation of the estimates and standard errors. The most significant limitation on accurate estimation is usually the sample size at the highest level (Maas & Hox, 2005). For instance, when a relatively large portion of the variance is situated at the therapist level, and randomization takes place at this level, power does not increase much when more patients are included per therapist. Similarly, without

prolonging the duration of the study, the effect of additional measurements on power is limited (Moerbeek, 2008; Raudenbusch, 1988). However, having a larger number of measurements does provide more information about the progress of individual patients and makes it possible to fit models with a more complicated random part at the patient level (Snijders & Bosker, 1999).

Covariates

The effect of covariates on power is not always straightforward in multilevel models. The optimal sample size for each level can change when covariates are used, particularly when covariates are used at multiple levels. Determining the precise impact of a covariate on power a priori is complicated because it depends on a number of factors, including how much within and between variance a covariate accounts for (Reise & Duan, 2003). For example, should a covariate have a low correlation with the dependent variable, but a moderate to high correlation with other variables in the model (such as time or treatment condition) power is usually decreased by including the covariate in the model. However, if the covariate explains part of the variance that is unexplained by time or treatment condition, power is increased by its inclusion (Moerbeek, 2006).

Missing data and drop-out

One of the challenges of longitudinal research is the, almost inevitable, loss of data due to missingness or drop-out. Multilevel analysis is capable of handling most cases of missing data very well[1]. Missing data will, nonetheless, affect study power simply by the fact that fewer data points are available as a result. Moerbeek (2008) presented power plots for several drop-out patterns and showed that in studies with dropout, power especially decreases when the drop-outs are concentrated in the beginning of the study. Increasing the study duration can also have a negative effect on power, especially if the dropout is concentrated at the end of the study.

Data used as example

For a new study on the effect of providing feedback to therapists about their patients' progress, we sought to estimate how many therapists and patients would need to be included in order to obtain sufficient power to detect an effect. In the planned study, patients will be randomly assigned to either a feedback or control condition. In the feedback condition, therapists will receive charts by e-mail that indicate the patients progress on the Outcome Questionnaire (OQ-45; Lambert et al., 2003). Because the

variance components in the planned study are unknown, data from a routine outcome monitoring study were used to estimate the variance in the control group. The routine outcome monitoring (ROM) data consists of 1966 measurements of patient functioning (average of 3.22 measurements per patient), within 610 patients (average of 5.60 patients per therapist), who were treated by 109 therapists. Collection of the ROM data is an ongoing process and a portion of the patients in this dataset are still in treatment. Patients completed the OQ-45 at several moments during treatment. We used the log of the number of sessions as our time variable, a common choice in psychotherapy research that has been shown to linearize treatment progress. The intraclass correlation at the patient level[2] and therapist level were 0.75 and 0.18 respectively in the (empty) unconditional model. Several three-level multilevel models were fitted to the data and compared on the deviance values and Wald test for fixed effects. For explanatory purposes we used relatively simple models, even though there are numerous modeling options that might be relevant for this kind of data (e.g. anchoring, centering, transformations, additional covariates). Table 1 shows the results for a model with a random effect of (log) time at the patient level (Model A) and the effect of adding a covariate (gender) to the model (Model B). Adding a random effect of (log)time at level-two and three simultaneously demonstrated no significant improvement compared to Model A. Adding a random slope at the therapist level,

Table 1 Parameter estimates and standard errors for two multilevel models fitted to the routine outome monitoring data

	Parameter		*Model A*		*Model B*	
			Estimate	*SE*	*Estimate*	*SE*
Fixed effects						
Initial status, β_{0jk}	Intercept	δ_{000}	78.37	1.26	78.38	1.27
Rate of change, β_{1jk}	Slope	δ_{100}	-13.30	0.99	-10.56	1.64
	Gender	δ_{110}			-4.16	1.99
Variance components						
Level 1	Within-person		71.77	3.27	71.67	3.26
Level 2	Initial status		454.09	23.97	448.93	30.69
	Rate of change		187.80	29.57	183.72	29.25
Level 3	Initial status		48.61	19.40	49.53	19.61
	Rate of change					
Goodness of fit	Deviance			16123.5		16119.24
	Pseudo R^2					

SE = Standard error
Model A: random slope at the patient level
Model B: random slope at the patient level, gender as covariate

but not at the patient level, produced a significant improvement compared to a fixed effect of time, however, the model with a random slope at the patient level had a better goodness-of-fit.

Once we derived our best-fit model, we used the values from Model A as estimates for the variance and covariance components in the control condition of the planned study. The equations used to compute the power curves are described in Appendix A. A medium effect of 0.5 was expected for the interaction between time and treatment condition. For the power plots, the study duration was set at 21 weeks[3], with patients receiving one therapy session per week. The maximum number of therapists was set at 100. To demonstrate the effect of adding a covariate, we used the values from Model B. To show the effect of drop-out on power, we performed a simulation study using Splus. For each combination of number of therapists and drop-out pattern, 5000 data sets were simulated in which model parameters and standard errors were estimated. The observed power is reflected by the percentage of data sets in which the null hypothesis was not rejected; in other words, in percentage of data sets in which the effect of the feedback condition on the slope was not significant.

A priori power analysis using the example data

Figure 1 shows the estimated power curve for five measurements per patient, and two, four and eight patients per therapist. Since the slope variance at the therapist level did not significantly deviate from 0 in Model A, the therapist variance has no influence on the power to detect an effect (also see Appendix A), and consequently the figure applies to both randomization on the patient and the therapist level.

General guidelines state that a power level of 0.80 is considered adequate to detect treatment effects (Cohen, 2002). According to Figure 1, to obtain a power of 0.80 one should have either 22 therapists with eight patients each (176 patients in total), 43 therapists with four patients each (172 patients in total) or 85 therapists with two patients each (170 patients in total). In order to get unbiased estimations of the parameters and standard errors, a sufficient number of therapists is needed. In the simulation study by Maas & Hox (2005) for two levels, it was shown that with 30 groups the standard errors for the highest level were estimated about 15% too small; only with 50 groups were the chances of reliable estimation acceptable.

Optimal level of randomization

In the ROM data, the slope at the therapist level was relatively small and not significant. However, in some cases there might be a significant difference between therapists in how their patients progress over time. In that case, the level of randomization could

Figure 1 Estimated power for the planned study, no significant slope variance at the therapist level

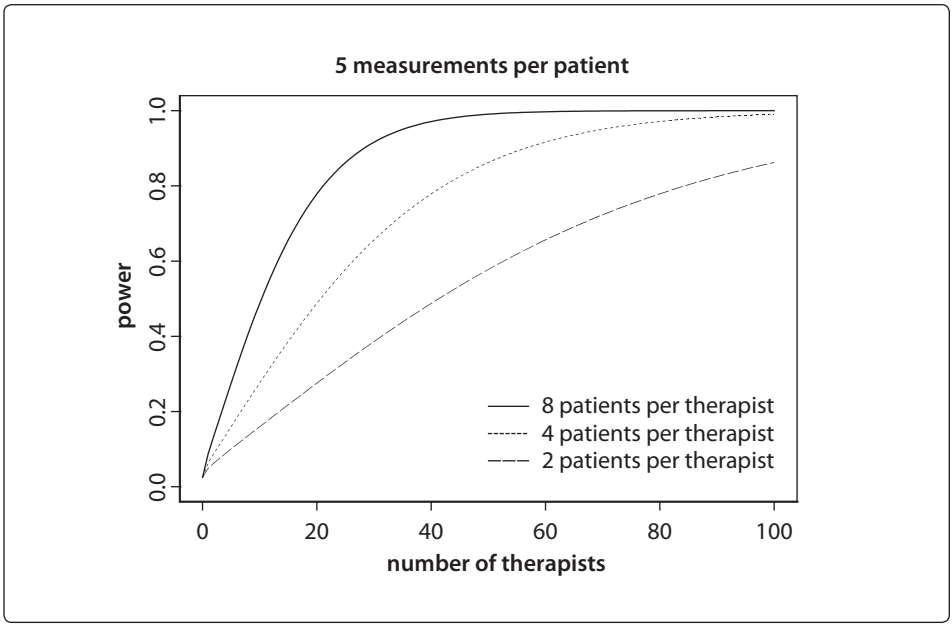
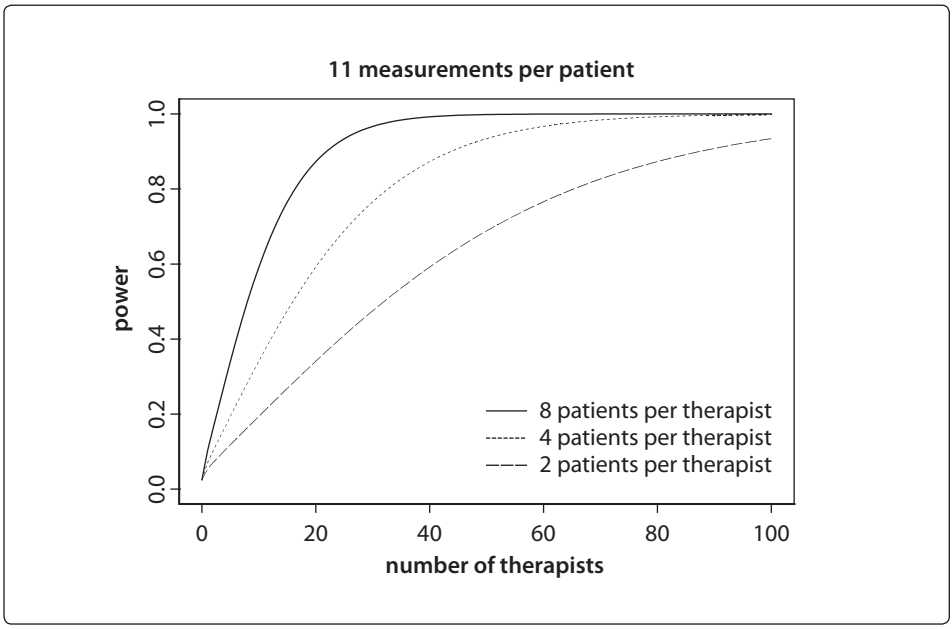


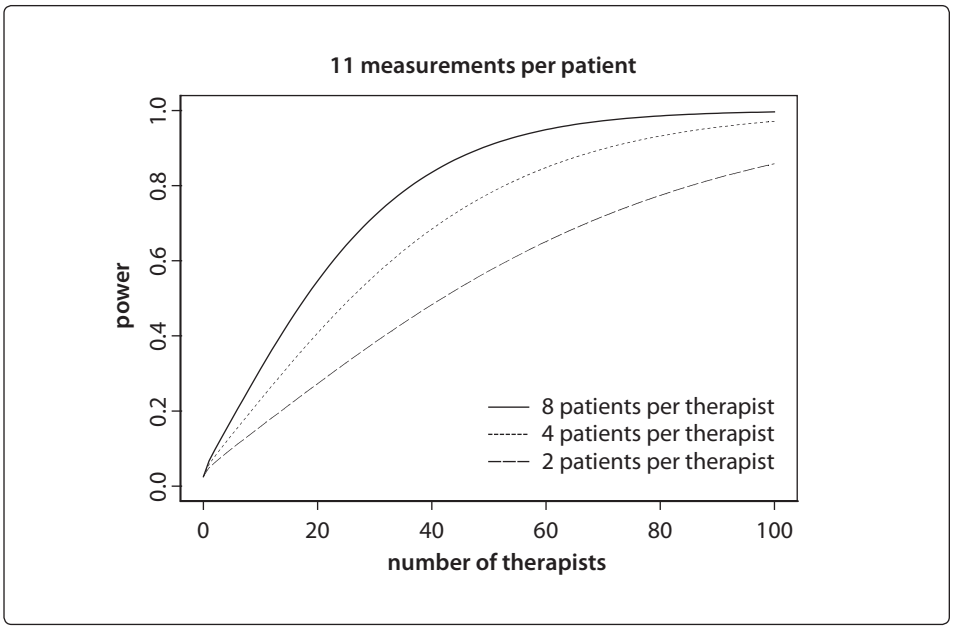
Figure 2a Estimated power, significant therapist slope variance, randomization at the patient level (level 2)



have a substantial impact on power. For this reason, we performed an additional power analysis with a significant slope at the therapist level. Based on the results from Lutz, Leon, Martinovich, Lyons & Stiles (2007), we assumed that beside the slope variance of 187.80 at the patient level that was found in Model A an additional variance of 35.77 (16% of the total slope variance) was situated at the therapist level. Figure 2a shows the power curve for this simulation, with randomization at the patient level, and Figure 2b shows the power curve for randomization at the therapist level, both for eight, four and two patients per therapist and 11 measurements per patient. When randomization takes place at the patient level (patients are randomly assigned to conditions, therapists are in both conditions) the samples required are 17 therapists with eight patients each, 33 therapists with four patients each, or 66 therapists with two patients each to reach a power level of 0.80. When randomization takes place at the therapist level (therapists are randomly assigned to conditions, all patients within a therapist are in the same condition) sample requirements are 37 therapists with eight patients each, 53 therapists with four patients each, or 86 therapists with two patients each.

Comparing Figures 2a and 2b also shows that adding patients per therapist has less effect on power when randomization takes place at the therapist level than at the patient level. Suppose we have 15 therapists with four patients each. When randomization occurs at the patient level, the power is 0.47. When the number of

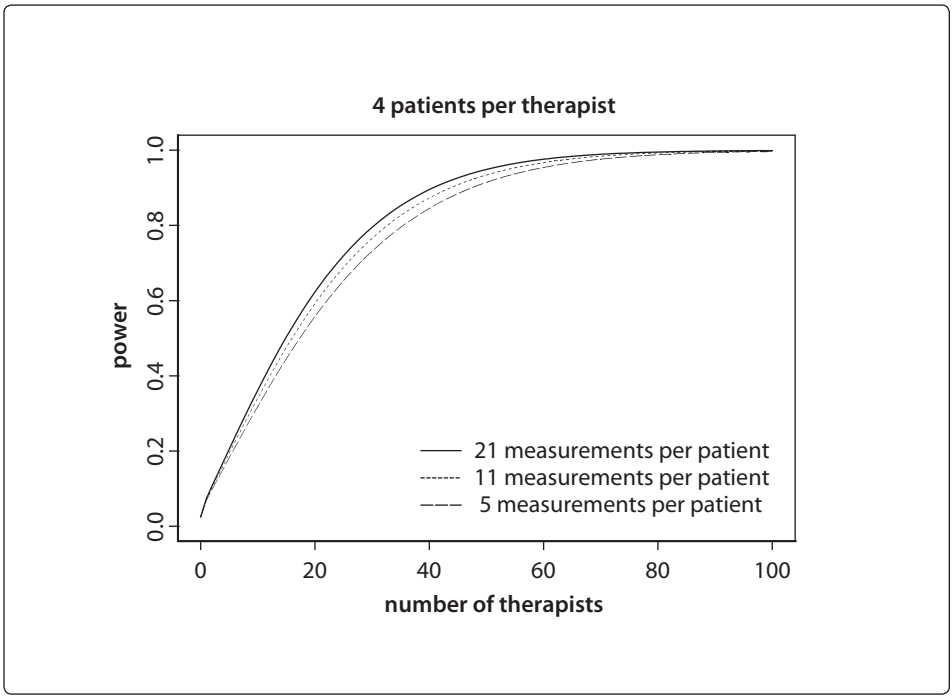
Figure 2b Estimated power, significant therapist slope variance, randomization at the therapist level (level 3)



patients per therapist is doubled, the resulting power is 0.77. When randomization takes place at the therapist level doubling the amount of patients per therapist increases the power from 0.32 to 0.44. However, doubling the amount of therapists in the study, rather than the amount of patients, results in a power of 0.56 and is, thus, far more effective.

Figures 2a and 2b demonstrate how randomization at the therapist level is less effective in terms of power than randomization at the patient level. Randomizing at the therapist level is only advised in case of contamination. In the planned study, contamination is not likely, as the studies performed by Lambert and colleagues (Harmon, Hawkins, Lambert, Slade, & Whipple, 2005; Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004; Lambert et al., 2002; Lambert et al., 2001; Whipple et al., 2003) show, therapists do not perform better with patients in a control condition over time. If contamination was an issue, effect sizes would have shrunk in each consecutive study as a result of the learning effect. In other words, if the feedback had taught therapists to treat all their patients more effectively, the effect of the feedback would have decreased with each consecutive study. Randomization at the patient level seems the best option for the study design in this case and, therefore, further power plots will only be presented for this situation.

Figure 3 Estimated power for a varying number of measurements per patient, randomization at the patient level



Number of measurement occasions

Earlier it was stated that the number of measurement occasions within the same time-period has little influence on power. To demonstrate this, power plots were produced for 5, 11 and 21 measurements. As can be seen in Figure 3 increasing the number of measurements per patient hardly has an effect on study power; respectively 36, 33 or 31 therapists are needed to obtain a power level of 0.80.

Adding a covariate

To demonstrate the effect of a covariate on power, gender was added as a slope predictor (see Figure 4). Although gender is a significant slope predictor, adding it to the model has little effect on power because it decreases the slope variance by only 2%. Stronger covariates might have a larger effect on power, provided that they don't increase variance on other levels. As is shown in Table 1, Model B, adding a covariate can increase the variance at other levels, in this case the initial status variance at level 3.

Figure 4 Estimated power with and without gender as a covariate, randomization at the patient level

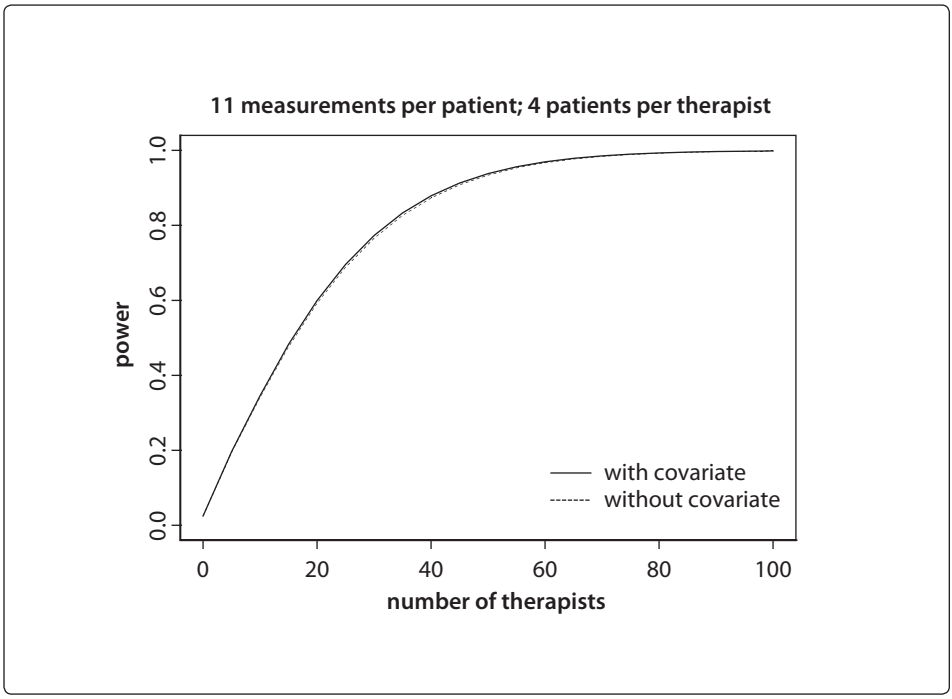


Figure 5a Simulated patterns of dropout (25% drop-out)

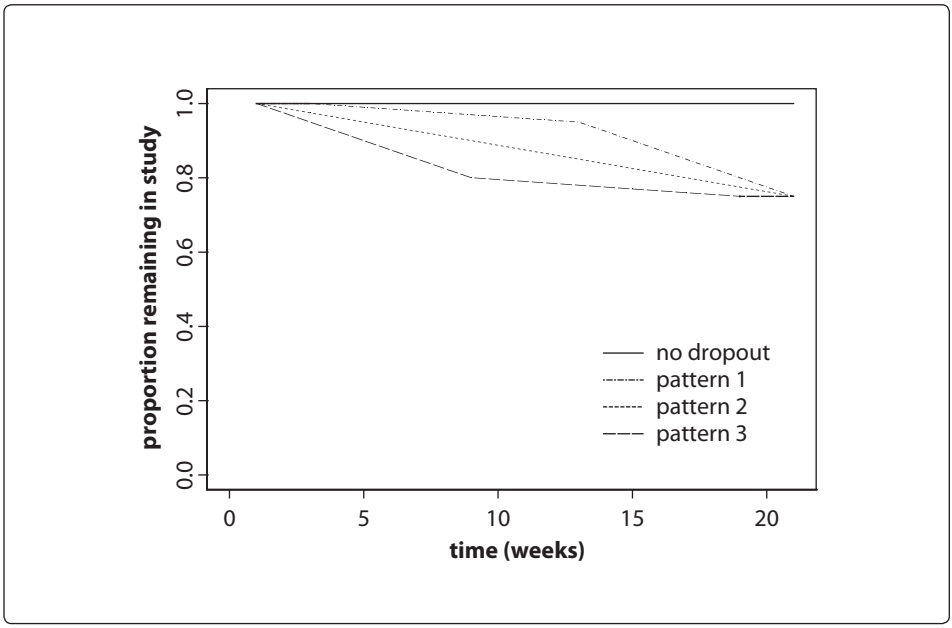
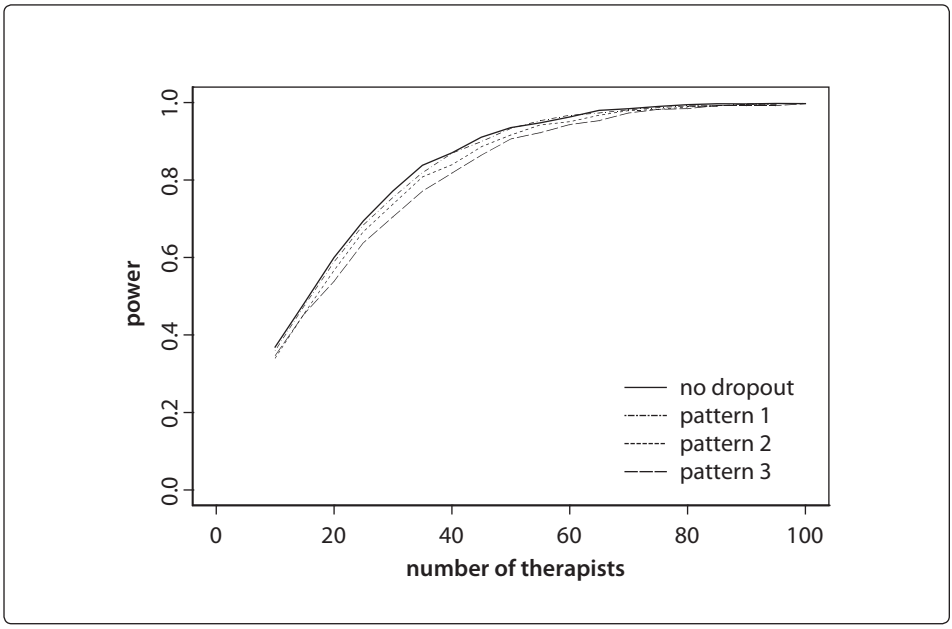


Figure 5b Estimated power for the simulated drop-out patterns, randomization at the patient level



The effect of study drop-out on power

Figure 5a shows the simulated patterns of drop-out for therapists with eleven measurements per patient and four patients per therapist. All patterns have a total drop-out of 25%. Pattern 1 has most of the drop-out at the end of the study; Pattern 2 has an equal drop-out throughout the study and Pattern 3 has most of the drop-out at the beginning of the study. Figure 5b shows the observed power for the simulated drop-out patterns. The effect of drop-out on the number of therapists needed for drop-out is limited. Having most of the drop-out at the beginning of the study seems to have the most substantial impact on power. Drop-out at the end of the study has the least impact on power. In a sample with no dropout, 31 therapists are needed. Drop-out at the end of the study results in a need for two additional therapists (33), equal drop-out throughout the study requires three additional therapists (34) and drop-out at the beginning in 6 additional therapists (37).

Discussion

The aim of this article was to perform an a priori power analysis for three-level longitudinal multilevel models and to demonstrate the effect of the level of randomization, samples size, covariates and drop-out on the power to detect a treatment effect. Results demonstrated that randomization at the patient level was more effective, in terms of power, than randomization at the therapist level. Increasing the number of patients was shown to be the best way to improve power when randomization takes place at the patient level. In the case of randomization at the therapist level, including more therapists in the study was more effective. The sample size at level one, the repeated measurements, did not have a strong effect on power. Furthermore, in our example adding gender as a covariate did not improve the power much. However, our covariate did not have a strong effect and other, more significant, covariates might have different effects on power. Drop-out also did not affect power substantially, although it did reduce power to some extent, especially when drop-out was concentrated at the beginning of the study. Besides power, it is necessary to have appropriate sample sizes at each level to ensure accurate estimation of parameters and standard errors. In some cases this may result in larger sample sizes than are necessary for sufficient power. In addition, in order to effectively distinguish between the slope variances at the patient and therapist level, there needs to be a sufficient level of patients per therapist.

Results indicate that, in three-level models, larger sample sizes are required than are common in general linear model approaches. It should be noted, however, that the intraclass correlation at level three in our example data was rather high.

This is consistent with naturalistic data, and one is less likely to find such values in a randomized controlled trial. Moreover, it is likely that the feedback condition in our planned study will reduce therapist variance and, by consequence, the level three intraclass correlation. In other naturalistic data we have found lower ICC values and, as a result, sufficient power with lower numbers. In such a case, power is no longer the main concern, and is secondary to estimation bias. For example, in our simulation, we found sufficient power using 17 therapists with eight patients each. However, by choosing that size sample, the model parameters and standard errors could be seriously biased, thus inflating the Type I error. This is specifically the case for the estimation of random effects; in fixed effects the standard errors are more robust. Maas and Hox (2005) found evidence for this in a simulated two-level model and although there are no simulations available for longitudinal three-level models, the problem of estimation bias likely applies to the highest level of the model and, thus, similar results can be expected for three-level models. Estimation bias in multilevel modeling is partly resolved by using non-parametric tests such as the likelihood ratio test instead of parametric tests, but this doesn't solve the problem completely.

The analyses that were performed in this article have some limitations. First, the modeling on which the analyses were based was kept deliberately simple. A random intercept, random slope model was selected, but fixed intercept or slope models could also have been used. For the time variable, a log-linear model was chosen, as this is the most frequently reported time variable in psychotherapy research, but other time variables are plausible as well (e.g. a combined linear and quadratic time variable). Another issue with the time variable is that calculating the dosage in sessions attended, rather than length of treatment, ignores frequency effects that would be found in, for example, twice weekly versus once-weekly psychotherapy. In addition, psychotherapy studies usually have several relevant predictor variables, rather than just one, and there would be predictors for both the slopes and the intercepts at the patient and therapist level. Since intercept predictors do not influence the power to detect a treatment effect, they were not included in the model. An additional limitation of this article is that it only describes power for situations with two conditions. Many clinical trials have three conditions, comparing two experimental conditions and one control condition. In such a case, the model becomes more complex quickly, because an additional (dummy) variable has to be added. One has to know in advance how the three conditions will affect slope and intercept variances. Lastly, although this article has provided examples how several relevant factors such as covariates and drop-out influence power, it has only addressed these phenomenon one at a time, whereas in practice multiple such factors frequently apply at the same time. However, these results offer a good indication of how each factor influences power in these models and could help in making decisions about study designs in the future.

Although a priori power analysis for multilevel models is a complex undertaking, these results indicate that such modeling is not only possible, but practical. Traditional power analyses for linear models does not distinguish between sample sizes at different levels and can lead to an underestimation of the number of cases that are needed. Therefore, although a growing literature on multi-level power analysis exists, further exploration is warranted. In particular, in order to be able to perform a priori power analyses for clinical trials, there is a strong need for more articles on variance components and intraclass correlations in different types of patients, treatments and outcome measures.

Since the special section in this journal on therapist effects, the discussion about whether therapist effects exist has become mixed up with the discussion on what models should be used to investigate them. Wampold & Bolt (2006) have stated that longitudinal models may increase patient variability and thereby reduce therapist effects. While this is true, patient variance in treatment progress is an integral part to the kind of data we collect in psychotherapy studies, as patients differ in their treatment course. By ignoring that variance one might be able to better detect therapist effects, but does that mean it is better? We do not claim to have the answer to that question, and would like to state that, in our opinion, it is more a matter of what the focus of the study is, than one method being better than the other. Longitudinal models do have some disadvantages, for instance, in the case of differential treatment length, longer treatments will have more impact on the model as they have more measurements than shorter treatments. Another issue is that longitudinal models do not assess whether treatment is successful or not, but neither do end-of-functioning models.

Irrespective of the method or model used, the evidence for therapist effects is limited so far (Crits-Christoph & Gallop, 2006). The lack of evidence for therapist effects in clinical trials seems to have two main causes: the number of trials that take the therapist level into account is small and the trials that do include therapist effects often have small sample sizes that are too small for unbiased estimation or sufficient power. In order to get more information on the existence of therapist effects in clinical trials, the therapist level should, at least, be registered and reported, regardless of significance. In addition, predictor variables at the third level should be included in more studies, to explain what factors may contribute to therapist effects.

Acknowledgements

Mirjam Moerbeek's research was funded by the Netherlands Organisation for Scientific Research (NWO), grant number 451-02-118. We would like to thank Sam Nordberg for the corrections he made to improve the English.

Notes

I Multilevel analysis can handle data that is Missing Completely At Random (MCAR) or Missing At Random (MAR), but not data that is Missing Not At Random (MNAR).

II The patient level variance has been calculated according to Davis & Scott, 1995 cited in Hox, 2002 (p.32)

III A study duration of 21 weeks results in equally spaced, round session numbers for the measurement occasions, as well as pre and post test. For 5 measurements the occasions are set at session 1, 6, 11, 16 and 21; for 11 measurements at sessions 1, 3, 5,…21; and for 21 measurements at each session.