Measuring and understanding patterns of change in intervention studies with children: implications for evidence based practice[1]

Julie E. Dockrell

Psychology and Human Development

Institute of Education

20 Bedford Way

London WC1A 0AL

E:j.dockrell@ioe.ac.uk

James Law

Centre for Integrated Healthcare Research

Queen Margaret University College

Edinburgh EH12 8TS

E: jlaw@qmuc.ac.uk

[1] We would like to thank ICAN , the health trusts involved and the two research officers, Kerry Williams and Belinda Seeff, who collected the data.

Abstract

Purpose

Comparisons across studies of the effects of intervention are problematic. Such analyses raise both methodological and statistical challenges. A single data set was examined to investigate whether different established approaches to measuring change in children with specific language impairments alter the conclusions that can be drawn regarding the efficacy of an intervention.

Methods

Measures of cognitive and language skills were collected at baseline and at six months following an intervention. Reliable and valid psychometric measures were used. Data from the intervention study were used to explore the patterns of results obtained using four different measures of change: change of diagnostic category, differential improvement across assessment measures, item specific changes and predictors of individual change.

Results

Associations between different tests purporting to measure similar constructs were modest. The measures identified different children as impaired both at baseline and follow-up. No effect of intervention was evident when a categorical analysis of impairment was used. Both treatment and comparison children changed significantly across time on the majority of measures, providing evidence of development, but specific effects of the intensive intervention were evident using ANCOVAs. Item analysis indicated that one of the standardized language tests adopted in the evaluation was insensitive to change over a six month period. Change in individual children's performance was predicted by language level on entry to the project.

Conclusion

The implications of the results are discussed in terms of the range of analytic approaches available to intervention researchers and the need to consider combinations of methods when analyzing outcome data.

Introduction

There is a major drive within health and education to provide evidence based services to children. To develop evidence based practice it is necessary to identify the specific effects of interventions. Decisions need to be made about the nature of the intervention, the measures used to examine change and the appropriate ways of analysing data. Using a single data set we consider the extent to which different methods of conceptualising and measuring change lead to different conclusions about developmental trajectories and the efficacy of different interventions. Data from young children with specific language impairments are used to illustrate the ways in which different analytical approaches can alter interpretations of the efficacy of interventions. Such comparisons should inform our understanding of analytic techniques used to evaluate change and the differential effects of interventions.

Measuring and examining change in children's language performance is important for the evaluation of interventions (Law, Garrett & Nye, 2003), for service development (Law, Lindsay, Peacey et al., 2001) and resource planning (Law, Dockrell, Castelnuevo et al., 2006). In addition such analyses have implications for our understanding of developmental pathways (Conti-Ramsden & Botting, 1999; Conti-Ramsden, Botting, Simkin et al., 2001; Law, Tomblin, & Zhang, submitted). This seemingly straightforward activity is fraught with complications. Decisions need to be made about the tools for assessment, the nature of the intervention and the way in which change in the population is measured. Our ability to draw reliable and valid conclusions about developmental trajectories is influenced by all of these factors (Zhang & Tomblin, 2003).

Using data collected from actual evaluations, as opposed to performing statistical modelling (see Wright, 2005), provides the opportunity to consider the

variation in effects at both the general (group) and the specific (the child) level. This provides the opportunity to compare different approaches using the same measures. Typically, single-subject experimental designs focus on targeted measures of language and do not evaluate their efficacy in terms of standardized measures while larger group interventions tend not to consider performance at the individual level (Bishop, Adams & Rosen, 2006). While both approaches have the potential to inform theoretical models the former speaks most directly to the practitioner, the latter to conventional models of deriving evidence based practice (Irwig, Glasziou & March, 1995). In this research note we consider four different approaches to examining change: (1) change of diagnostic category implicating service eligibility, (2) differential improvement in language measures, (3) item specific changes on tests, and (4) predictors of individual change in test scores.

At a practical level, change can be evaluated in terms of whether children can be considered as eligible or not for support services. All things remaining equal change in eligibility for speech and language service following an intervention implies that the intervention was effective. Thus, it is possible to examine changes in service need over time as an indicator of change in language competence. For children with SLI, and other learning needs, service eligibility is often made on the basis of the relationship between cognitive ability and assessments of performance on a target variable such as language, reading or numeracy. For children with SLI this typically entails a significant discrepancy between their language skills and non verbal ability (Dockrell, Lindsay, Letchford & Mackie, 2006) despite the fact that the view that cognition sets the upper limit for development is open to dispute (Cole, Dale, & Mills, 1992; Cole, Schwartz, Notari et al, 1995). A discrepancy between language and non-verbal ability is the most common criterion to define eligibility for speech and

language services for specific language difficulties in the UK (Dockrell et al., 2006). Changes that reduce this discrepancy have implications for future service provision and expectation of the children's level of need. Thus, one approach to measure change is to examine changes in children classified as experiencing a language problem and thereby requiring targeted intervention. There are, however, both conceptual and statistical reasons why categorisation may not be clinically appropriate (Botting, 2005; Wright, 2003).

An alternative and commonly used approach is to focus at the group level and examine changes in language scores as a result of the intervention (see Matheny & Panagos, 1978; Girolametto, Pearce & Weitzman, 1996a; Robertson & Weismer, 1999; Bishop et al, 2006). For the majority of children some change will occur in test performance in the absence of any targeted intervention (Dockrell, Stuart & King, 2006). To prevent erroneous conclusions being drawn it is therefore essential to include control or comparison groups (Cohen et al., 2005). Comparison groups provide the added advantage of a cohort that has experienced some intervention thus, allowing the measurement of differential patterns of change and addressing the statistical confound of regression to the mean. However, typically it is often not possible to allocate participants randomly to separate interventions (Wertz, 2002) and evaluation of change needs to control for differences in baseline measures. A simple comparison of post-treatment results is unwise since it confounds potential to change with actual change. Such difficulties can be minimized by use of analyses that control for this by considering gain scores (normalised gain score, Hake, 1998; Ebbels, Van der Lely, & Dockrell, in press). Gain scores can be analysed in at least three different ways: *t* tests based on gain scores, ANCOVA or split plot. Some statisticians argue that "in most cases you should analyse the data in several ways" (Wright, 2003; 130)

to explore the patterns within the data set. These comparisons are virtually never published outside the statistical literature (but see Wright, 2006). The most rigorous approach is an ANCOVA controlling for initial performance.

Group analyses of total test scores have been criticized because they fail to differentiate individual item responsiveness within tests (Prieler & Raven, 2002). For example, equal raw score changes between low and high language performers do not necessarily imply equal differences in language competence. Potential changes in scores are influenced both by child performance level on a test and by the test properties. As such it can be difficult to draw valid conclusions about the relative gains of high and low ability children in response to intervention. It is possible to control for such limitations by examining participants' changes using single item changes (Prieler, 2000). These analyses allow an evaluation of both the test and the child.

An important consideration in the evaluation of change is the measurement tools designed to assess competence (Dockrell, 2001; Stuart & Stainthorp, 2004). There is concern about the numerous and diverse range of instruments that are used to identify children as having language impairments (McCauley & Demetras, 1990). Indeed the use of these measures can lead to quite different profiles of performance. An early study by Howlin and Cross (1994) highlighted this point. They demonstrated how children apparently developing normally provided quite different profiles on measures, which, although different, ostensibly measure the same construct. They tested children on six well standardized and commonly used language measures including the Test of Reception of Grammar (TROG, Bishop, 1983), Reynell Developmental Language Scales (Reynell & Huntley, 1985), British Picture Vocabulary Scales (BPVS, Dunn et al., 1997) and the Bus Story and Action Picture

Test of the Renfrew Language Scales (Renfrew, 1978). The results revealed that, while on some measures the children's scores were within the normal range for their chronological age, on other measures they showed a marked discrepancy for their chronological age. Although there have been improvements in the measures used to assess language development since this study, measurement error remains a significant limitation. Even the most reliable and valid standardized tests used to assess children's language skills have limitations (McCauley & Swisher, 1984; McCauley & Swisher, 1986). It has been argued that use of standardised tests can lead to the identification of normal children as language impaired, the provision of misleading profiles of verbal and nonverbal performance, an inability to estimate the severity or describe the general nature of the language impairment and, result in an increase in the number of children identified as language impaired with each successive re-norming of the measures concerned (McFadden, 1996). To characterise the nature and differences across measures, comparative test performance is needed; preferably, initially, on a typically developing sample (see Nation & Snowling, 1997 for this approach with reading). In general these data are not available for language tests.

The data used for the current analyses were collected as part of an evaluation of preschool provision developed to improve the language skills of children identified as experiencing SLI (see Law, Dockrell, Williams, & Seeff, 2001). Following Cole, Dale and Mills (1992) language eligibility consisted of performance one *SD* or more below the mean on the criterion language assessment. This criterion represented the maximum language performance for inclusion. The target intervention was offered in Early Years Centres (EYCs) for children with SLI. The model, which is not the focus of this paper, is characterised as a short term intensive form of service delivery arising

out of a close collaboration of educational and speech and language therapy services (Law et al, 2001). The comparison intervention was "routine" clinical practice in the UK (Glogowska, Roulstone, Enderby, & Peters, 2000). This paper explores ways in which research questions inform the data analysis and impact on the interpretation of the efficacy of interventions. By using the same data set for the analyses we are able to consider the strengths and limitations of the different analytical approaches. We considered four different ways of evaluating children's language performance and change: (a) change in diagnostic category over time for all children and across interventions; (b) change in performance over time on standard scores for all children and across interventions; (c) change in performance over time on individual test items for all children and across interventions; and (d) predictors of change.

Methods

Participants

A total of 91 children participated in the study. Children with significant speech and language delays associated with intellectual disabilities or those with other physical impairments were not included in the study. All the children were identified by speech and language therapists as having SLI, that is difficulties that were not associated with other known conditions, and all children had a nonverbal IQ score within 1 *SD* from the mean in conjunction with significant language delays (Law, Dockrell, Williams, & Seeff, 2001). There were 58 children (45 boys and 13 girls) in the EYC intervention group with a mean age of 40 months, *SD* = 6, range 33-53. The majority (87.9%) had English as their first language. These children were matched by age and non-verbal ability with thirty-three children (20 boys and 13 girls) with a mean age of 42 months, *SD* = 6, range = 32-58 who formed a comparison intervention group. The comparison children were drawn from adjacent health services with

comparable demographics to the EYC intervention group. There were no significant difference between the groups in age at time of entry into the evaluation, F (1, 89) = .746, *ns*) nor on any language or cognitive measure (see results). All children were assessed immediately prior to the programme, immediately after the intervention period when a language sample was collected and parents interviewed and then again six months after the initial assessment when standardized measures were readministered. The data presented here reflect the first and the last of these data collection points.

Assessment measures

Rationale for test choice

Evaluation of language competence and change is commonly assessed through the use of standardized tests especially a composite measure of language (see Law et al., 1998 and Law et al., 2003 for a summary of assessments used in interventions studies) although overall evidence about diagnostic or predictive properties addressing language is 'weak and incomplete at this time' (Health Service Technology Assessment, HSTAT, 2006). In clinical practice psychometric adequacy is not necessarily the determining factor in test choice; often tests are used for more pragmatic reasons (Huang, Hopkins & Nippold, 1997).

It is, however, possible to limit the potential problems associated with standardized language measures. Tests should be based on an appropriate standardization sample, and therefore provide a reliable measure of a child's relative standing in comparison to developmental language norms. Thus, our first criterion for choice of assessment measures was that they were contemporary and met high standards of reliability and validity. A second factor in our choice of measures was that they were in common usage by practitioners in the field to evaluate expressive

and receptive language, thereby providing clinically relevant data. Finally, we wanted to use one test that included both a measure of language and a measure of non-verbal skills so as to reduce testing error in using different instruments to develop a performance profile. Our criteria led us to two measures, both with UK norms: Preschool Language Scale UK (PLS-3) (Zimmerman, Steiner & Pond 1992) and the British Abilities Scale II (BAS II Elliot, Smith & McCulloch, 1997). The PLS-3(UK) is commonly used in clinical practice and is the only pre-school language measure to meet acceptable criteria for validity, normative data and the relaxed criteria of reliability for language assessment measures (HSTAT, 2006).

Like the PLS3 UK, the British Abilities Scale II (BAS) benefits from recent re-standardisation with a representative sample of the population (Elliot & McCulloch, 1997). High levels of reliability and validity are reported. Since the assessment aims to map information processing systems to psychometric assessment profiles it addresses some of the significant limitations of previous assessment measures (Hill, 2005). This is an early year's scale, which allows a separation into verbal ability (Naming and Comprehension) and non-verbal ability (Picture similarities and Block building). The test is used extensively both for research with preschool populations and in the practice of educational and clinical psychologists (Hill, 2005).Only tests which had internal consistency of .8 and above and test retest correlation of .8 were included in the battery. Two different assessment measures were used to profile the children's strengths and needs.

The *British Ability Scales II* (BAS II, Elliot, Smith, & McCulloch, 1997) was used to assess the children's nonverbal and verbal abilities. The five age appropriate subtests that were administered were Block Building, Picture Similarities, Verbal Comprehension, Naming Vocabulary and Early Number Concepts. Scores are

presented as T-Scores and Ability Scores. Block Building and Picture Similarities combine to provide a composite measure of *nonverbal ability* and Verbal Comprehension and Naming Vocabulary combine to provide a composite measure of *verbal ability*. The test has been appropriately standardised on a British population.

Language skills were examined using the *Pre-school Language Scale-3* (PLS-3 (UK); Zimmerman, Steiner, & Pond, 1992) to examine language comprehension and expressive language. The concurrent validity of the PLS-3 with other standardised measures of language includes 0.52 with the Test of Early Language Development for typically developing 4 year olds. The coefficient for the BPVS was lower for typically developing children (0.29) but higher (0.59) for children enrolled in Head Start programmes. The comparable figures for ability measures include 0.55 for the correlation between the Kaufman Assessment Battery for Children and the PLS-3 auditory comprehension score (Zimmerman, Steiner & Pond, 1992).

The nature of the interventions

Early Years Centres. The EYCs provided intensive, multi-professional support for children with identified speech and language needs over relatively short periods of time (six to ten weeks for 2 hours and 30 min a day). The staffing in the centres included teachers and nursery support staff, speech and language therapists and educational psychologists. Interventions include a structured language curriculum with individualised planning and daily intervention. A description of the programmes of the centres is presented elsewhere (Law et al., 2001; Law et al., 2005).

Typical therapy. The intervention received by the comparison group was made up of typical provision in health service settings within the UK (Glogowska et al., 2000; Law & Conti-Ramsden, 2000). Characteristically the child would be seen

individually with the parent but might subsequently be included in therapeutic groups. Children were seen within the child's local health centre and there was no explicit attempt to link the therapeutic activity with wider nursery school objectives. Children in the comparison group varied considerably in the contact they had with the speech and language therapy services over the 6 month time frame (M = 4.7 hours, SD = 6.7). Some experienced regular individual therapy but for the many contact was intermittent.

Procedure

Prior to commencing the study, permission was obtained from health trusts, nurseries and parents for the children to participate in the study. The study was passed through local ethical procedures and all parents agreed for their children to participate in the study. All children attending the EYC centres participated in the project. The children were assessed at home on all test measures by an experienced speech and language therapist and a psychologist prior to beginning the intervention. In each case the speech and language therapist completed the speech and language measures and the BAS II was completed by the psychologist. Assessors were not involved in the implementation of the intervention nor were they aware of the individual children's specific intervention programmes. They were also blind to each other's assessment results at each phase of the study.

The comparison group was selected from cases seen routinely by local health service providers. Speech and language therapists were requested to refer any child who was on their waiting list or had been assessed but not begun treatment, and fulfilled the following criteria: age between 2.7 to 4.4 years; receptive or expressive language difficulties without speech difficulties; the child's speech or language

difficulty was not thought to be a result of any known cause such as cerebral palsy or sensori-neural hearing loss. Children classified as being on the autistic spectrum or with nonverbal scores greater than 1 *SD* below the mean were excluded. Identified children were also all assessed at home by an experienced therapist and a psychologist prior to beginning therapy.

All children were re-assessed on both measures six months after the date of first assessment. Assessors were, again, blind to each other's assessments.

Results

Table 1 presents the cognitive and language scores of the children on entry to the programmes. To allow an initial comparison across the different measures test scores have been transformed to Z scores. As expected all language scores differed significantly from the expected pattern for a typically developing population. A comparison was made of the verbal and nonverbal scores of the groups. Overall the children performed significantly better on the nonverbal tests than on the verbal tests of the BAS-II, F (1, 90) = 15.136, $p < .001$, $\eta p^2 = .16$). Thus, as a whole, the identified sample reflected a group of children who were experiencing specific difficulties with language but for many of the children, as Table 1 shows, this was associated with commensurate difficulties with numeracy and reduced performance on tests of nonverbal ability. A series of ANOVAs were used to examine the differential pattern of performance at baseline between the EYC and the comparison groups. The groups did not differ significantly at baseline on any of the language measures (PLS-3(UK) auditory comprehension standard score, $F(1,90) = 1.414$, $p = .238$; PLS-3(UK) expressive language standard score, $F(1,90) = 0.007$, $p = .934$; BAS-II verbal ability, $F(1,90) = 0.968$, $p = .328$); BAS-II nonverbal ability, $F(1,90) = .719$, $p = .399$; Early

number concepts, $F(1,90) = 0.062$, $p = .803$). Thus, as a group, the children met the conventional criteria for a specific language difficulty.

A series of correlations using standard scores examined the relationship between the language measures at Time 1. As Table 2 shows, all language measures were positively associated (at the .001 level) thus meeting Bonferonni correction levels of .01.

INSERT TABLE 1 AND 2 ABOUT HERE

The different tests, while having a degree of convergence, lead to different identification rates when performance of more than 1SD below the mean was used as a cut-off for the identification of a language difficulty. Thus, we initially considered an analysis that classified children in terms of specific language difficulties across the different language measures. To identify a clinically significant delay two different diagnostic groups were established: children where performance on both PLS-3(UK) scales was below 1*SD* (Low PLS) children where both BAS-II language measures were below 1*SD* (Low BAS). Overall 47 (51.6%) children met the criterion for PLS problem while 43 (47.3%) children met the criterion for BAS problem on entry to the study. They did not differ by intervention group (Time 1: PLS problem $X^2 = .67$, $df = 1$, $p = .796$; BAS problem $X^2 = 1.764$, $df = 1$, $p = .184$).

At follow-up (Time 2) 55 (70%) of the children were identified as having a problem on the PLS diagnostic criterion and 32% (29) on the BAS diagnostic criterion. The distributions did not differ across the intervention and comparison groups for either measure at the follow up (Time 2: PLS problem $X^2 = 1.725$, $df = 1$, $p = .189$; BAS problem ($X^2 = .51$, $df = 1$, $p = .821$). Thus change in category

identification would indicate no significant effect of the two interventions, as there were no follow-up differences between the two groups although trends for an increase in the BAS scores and a decrease in the PLS scores were evident.

However, the measures did not identify the same children at Time 1 or Time 2. Eight children (9%) at Time 1 identified as not having a problem on the PLS were identified as having a problem on the BAS and 12 children (13%) identified as not having a problem as identified on the BAS were identified as having a problem on the PLS. Patterns of identification differed significantly (X^2 =28.886, df = 1, p < .001). Thus, at Time 1, 22% of the sample received different classifications on the basis of the two tests and these figures were larger when subtest comparisons were used (see Law, et al., 2001). At Time 2 twenty-seven children (30%) were identified as having a problem on the PLS but not on the BAS and one child was identified as having a problem on the BAS but not on the PLS. Since both tests report good measures of reliability and validity in their construction such categorisation differences raise important questions about our understanding of population parameters and change over time. Our next analysis explored relative gains across time and measures to provide greater discrimination of development and change.

Standard scores were available for all measures of language and cognitive skills. Children's attainments at baseline and follow-up on all standardised cognitive measures and verbal measures are presented in Table 3 for the BAS-II and Table 4 for the PLS.

INSERT TABLE 3 AND 4 ABOUT HERE

To examine children's improvement relative to their performance at Time 1 a series of ANCOVAS[2] were carried out on the children's standard gain scores, thus overall gain was examined while adjusting Time 2 scores for Time 1 variability. It was not possible to analyse data from the PLS-3(UK) expressive measure due to the skewed nature of the gain scores ($Z = 1.436$, $p = .032$).

There was a statistically significant relationship between all baseline measurements and gain scores (BAS-II early number, $F(1,90) = 30.182$, $p < .001$, $\eta p^2 = .26$; BAS-II block building, $F(1,90) = 25.894$, $p < .001$ $\eta p^2 = .26$; BAS-II picture similarities, $F(1,90) = 58.817$, $p < .001$, $\eta p^2 = .40$; BAS-II comprehension, $F(1,90) = 47.503$, $p < .001$, $\eta p^2 = .35$; BAS-II naming, $F(1,90) = 42.408$, $p < .001$, $\eta p^2 = .33$; and the PLS-3(UK) auditory score, $F(1,90) = 12.163$, $p = .001$, $\eta p^2 = .12$). Thus these data indicated that both the children's cognitive and language skills improved over time in terms of standard scores.

No effect of group was detected for the PLS-3(UK) auditory, BAS-II early number, BAS-II block building, picture similarities or BAS-II naming; that is the mean change for the EYC group was comparable to that of the comparison group with the same baseline. However, this was not the case for the BAS-II comprehension measure, where there was a detected group difference, $F(1, 90) = 5.702$, $p = .019$ $\eta p^2 = .06$) with children in the EYC group improving on average 3.2 T score points more that the comparison group. These results were tested using a normalized gain score for comprehension confirming a significant effect of group on the measure, $F(1, 90) = 9.639$, $p = .003$, $\eta p^2 = .10$).

[2] ANCOVA is generally the preferred methods of analysis for interventions of this type. However, a series of repeated ANOVAS on the same data set provided the same results.

INSERT FIGURE 1 ABOUT HERE

Thus, on average, the BAS-II comprehension scores of the EYC group increased more than the comparison group relative to baseline comprehension scores. These data suggest that improvement on this measure was intervention related and is unlikely to be explained by regression to the mean.

The failure to demonstrate improvement on the other measures could reflect an intervention effect but equally the test may provide insufficient items to demonstrate change or items which are not sufficiently sensitive to the changes observed. To examine test sensitivity a linear logistic model with relaxed assumptions was used to examine children's success on individual items across time. This method does not require Rasch homogeneous data so it was possible to consider both the BAS-II language measure raw scores and the PLS-3(UK) measures. The analysis examined only individual items that changed from Time 1 to Time 2. Items from the BAS-II Naming Vocabulary scale provided sufficient data for the analysis and this resulted in a significant effect of time (Effect parameter 1.2155, SE = 0.4869, Z = 2.4965, $p <$ .05) but no effect of intervention group (Effect parameter –0.0444, SE = 0.9953, Z = 0.1502, $p = ns$). The BAS-II Comprehension scale also provided sufficient change items to assess change for these data. Again both groups had a positive significant change in performance over time (Effect parameter 1.2155, SE = 0.4869, Z = 2.4965, $p < .05$) and, in this case, the change in the children in the intervention group was significantly different from that of the comparison group (Effect parameter 1.2528, SE = 0.4604, Z = 2.7207, $p < .01$). Data from the PLS-3(UK) provided very few items that changed over the time, seven for the auditory scale and two for the expressive

scale thus resulting in a weak finding since so many items are dropped from the analysis. Again, there was an effect of time (Auditory Effect parameter 1.4916 *SE* = 0.3689, *Z* = 4.0435, $p < .01$; Expressive Effect parameter 2.0149, *SE* = 0.7520, *Z* = 2.6793, $p < .01$) but there was no effect of group in either case (Auditory Effect parameter 0.1638 *SE* = 0.4538, *Z* = 0.369, *ns*; Expressive Effect parameter –1.2041 *SE* = 0.9595, *Z* = 1.2549, *ns*). These results complement those described for the gain score analyses where the children in the intervention group demonstrated a differential positive effect on BAS-II Comprehension but not Naming. In addition the analysis explains the limited results for the PLS-3(UK); there were simply insufficient items to demonstrate change in the cohort.

The final analysis examined whether it was possible to predict which children changed over the course of the intervention. The focus here was on the area of receptive language as measured by the BAS because this was where significant differences were detected between the groups in the earlier analyses. Moreover since ability scores reflect item difficulty, are not dependent on norm referencing and show good discrimination we used these scores as our indicator of change. Of particular interest is the comparison between those children whose ability scores do change and those whose scores remain the same or decline over the period.

Of the 91 children in the study 51 of the 58 children in the EYC group (88%) and 20 of the 33 children in the comparison group (61%), a total of 71 (78%) had ability scores which changed in a positive direction over the course of the study. An ability score is an indication of the level of item difficulty that the child can complete successfully, it is a criterion-referenced score. The test performance of the two groups, those that changed in terms of their standardised scores and those that did not are given in Table 5. A logistic regression analysis was employed to test whether it was

possible to predict which children had ability scores that changed (improved over time) and those whose scores declined or remained the same (non-improver). The dependent binary variable was improver versus non improver on the children's ability score on the verbal comprehension scale of the BAS-II. The independent variables entered into the analysis were the children's age, gender, hours of therapy, whether they had or had not been in the EYC intervention group, their block building skills (ability score) and their receptive language skills as measured by the ability score on the BAS-II at baseline (i.e. before the start of the intervention). The results of the analysis indicated that the main factor to contribute significantly to the variance was the children's initial receptive language score on the BAS-II (B, -.048, SE 0.18, 6.925 p = .009). None of the other variables entered into the analysis were statistically significant. The group that did change had lower BAS-II block building ability scores (with mean 59.21 compared to mean of 87.25) were on average slightly younger than those that did not change, and had received substantially more speech and language therapy and other input (60 hours compared to 28 hours). Children with the poorer language comprehension scores as measured by ability scores on the BAS-II were most likely to improve over the course of the two interventions.

Discussion

Four research questions were framed at the outset of this paper and each will be addressed in turn. The wider implications will then be discussed. The data from this study suggest that the assessments used were related to a statistically significant degree but for measures purporting to assess parallel behaviours there was a substantial amount of variance that was unaccounted for.

A categorical analysis describing children as impaired (or not) revealed differences both at Time 1, when children were identified, and at Time 2, when the follow-up assessments were completed. Indeed at Time 2 32% were differentially identified by these two 'valid and reliable measures'. No differential effect of treatment was evident in these analyses. In contrast, analyses controlling for Time 1 scores and examining the extent of the children's progress revealed improvement over Time on the majority of standardised measures. Overall positive change on these measures may best be explained as regression to the mean. However, an intervention specific effect was evident for verbal comprehension measured by the BAS-II, but the effect is small accounting for only 6% of the variance.

To examine the extent test properties were responsible for the children's changes in performance we examined the potential for change in test items. This is a novel form of analysis for interventions in general and language interventions specifically. The current results demonstrate its utility for confirming differences from more standard analyses and in revealing test limitations for measuring change. Importantly there were too few items on the PLS-3(UK) that changed over a six month period and children were too inconsistent in their performance on those measures that there were available to identify change to describe either development or intervention effects. This analysis provided confirmatory evidence of the specific intervention effect for comprehension measured by the BAS-II. Our final analysis examined the profiles of children who changed or failed to change on the verbal comprehension scale of the BAS; the measure shown to be sensitive to differential change in the current study. The logistic regression indicated that initial receptive vocabulary level was the only significant predictor of whether children changed.

The general discrepancies between different analytical techniques raise fundamental questions about the inferences that are drawn about the identification and classification of children with specific difficulties. The data support earlier criticisms of the use of cut off points and emphasize the importance of considering the change in individuals scores rather than an attempt to classify the child as impaired or not. Diagnostic categories, in this context, speak neither to the child's level of need nor to the efficacy of the intervention. Rather they highlight the importance of considering response to intervention for individual children (Justice, 2006). Response to intervention is premised on the use of appropriate assessment tools and evidenced-based interventions.

Despite significant correlations between tests these vary between tests and across time. Moreover our third analysis demonstrates that even when tests are reportedly psychometrically robust they may be insensitive to developmental change and therefore inappropriate measures of intervention effects. To date, reliance has been placed on the identification of SLI but if reliable and valid measures perform as differently over time as they have done in the present study the validity of such claims is questionable. This has implications for researchers attempting to identify characteristic features of language-impaired populations.

Co varying for initial language scores provided a means of assessing intervention specific effects by group thereby allowing discrimination between measures and across interventions. These analyses identified an intervention effect for comprehension. However the analysis by improvers and non-improvers on ability scores raises important caveats to this conclusion. These data suggest that, despite the scope for all children to change, it was those with the poorer language competencies that improved. The improvement for both cohorts on the majority of standardised

measures might reflect relative improvement but in this context is more likely to reflect regression to the mean. The differential change in the children's ability comprehension scores, as measured by the improvers and non-improvers, is therefore an important result worthy of further evaluation. Data from existing intervention studies indicate that there is much less evidence about the effect of intervention on verbal comprehension than there is on expressive language skills (Law, Garrett, & Nye, 2003). Unlike the studies in the Law et al. (2000, 2003, 2005) review these children were not randomly allocated to intervention and control groups and the question remains whether this result is a function of initial selection bias or an assessment bias despite attempts to control for this methodologically and statistically. It is also possible that verbal comprehension may be more susceptible to change if the skills concerned are at an early developmental level. The data from the logistic regression would support this view. Thus it may be easier to shift a child whose comprehension skills are at a single word level to understand more single words or to understand two word phrases than it is to increase child's comprehension when complex grammatical forms are examined (Ebbels et al., in press). This may be a linguistic phenomenon, but it is also possible that the nature of the intervention group effectively targets listening and attention skills and the increase is reflected in the comprehension measure. We predict that interventions in other areas of development will experience similar problems of interpretation.

We began by questioning the ways in which interventions can be evaluated. The data presented here indicate that overall group improvements can be recorded on standardised measures even when a significant minority of participants fail to change their raw score performance over a six month period. This highlights the need to gain a greater understanding of what is a *typical* developmental trajectory. In addition

studies which evaluate interventions using standardised measures need to consider the inclusion of intervention specific measures, as are typically used in single-subject experimental design studies. The combination of both intervention specific measures and standardised tests scores should provide robust information about the validity of the specific interventions and allow an evidence based approach to service provision to be developed.

To provide reliable and valid information about the efficacy of interventions it is necessary to conduct systematic reviews (e.g., Law, Garrett, & Nye, 2003). Systematic reviews are at the heart of evidence-based practice; however, such analyses need to be based on studies with both robust methodology and appropriate statistical analysis. The conclusions drawn from such reviews are prefaced by the assumption that that the interventions and outcomes measures are sufficiently homogenous to warrant aggregation (Petticrew & Roberts, 2006). The responses to the four questions addressed in this paper demonstrate the need to look carefully at such studies and the ways that measures were employed to assess change in children following intervention. Authors conducting systematic reviews are encouraged to examine the trial quality of each included study. Our study indicates the need to incorporate issues related to sensitivity of measures as part of such a quality assessment.

Our data suggest that care needs to be taken when a single analytic technique is used to evaluate intervention effects. Evaluations of intervention in the area of language development and by implication development more generally need to consider both appropriate controls in the use of statistical methods and a systematic examination of the tool used to measure change. Careful use of these techniques can

provide relevant information about the efficacy of the intervention and profiles of those children who may benefit most from specific interventions.

Conclusions

To date emphasis has been placed either on establishing whether an intervention works for a given group of children or investigating the performance of individual subjects. This paper suggests that a typical field study of clinical effectiveness is able to provide considerable detail not only about children at a group level but also at an individual level and about the appropriateness of measures that are used. Importantly, item analysis can contribute to the understanding of whether specific measures are more valid to measure change in the characteristically noisy phenomenon such as early language development. Use of complimentary analytic approaches provides the basis for distinguishing between developmental effects, intervention effects and test factors.

Acknowledgements

The authors wish to acknowledge the contribution of the Department of Health in the UK for funding this project through I CAN, Kerry Williams and Belinda Seeff for their contribution to the data collection, Jorg Prieler for his contribution to the interpretation of the item analysis and Dan Wright for his statistical advice.

References

Beitchman, J. H., Wilson, B., Brownlie, E. B., Walters, H., & Lancee, W. (1996) Long term consistency in speech/language profiles 1: Developmental and Academic outcomes. *Journal of American Academy of Child Psychiatry*, *35*(6), 804-825.

Bishop, D.V.M. (1983). *Test of reception of Grammar*. Published by the author and available from Age and Cognitive Performance Research Centre, University of Manchester.

Bishop DVM, Adams CV, & Rosen S (2006). Resistance of grammatical impairment to computerized comprehension training in children with specific and non-specific language impairments. *International Journal of Language & Communication Disorders* 41 (1): 19-40.

Castelnuovo, E., Williams, K, Seeff, B., Dockrell, J.E., Law, J. & Normand, C. (2006). Early Years Centres services for pre-school children with primary language difficulties: what do they cost, and are they cost-effective? *International Journal of Language and Communication Disorders 41,* 67-81.

Conti-Ramsden, G., & Botting, N. (1999). Classification of children with specific language impairment: longitudinal considerations. *Journal of Speech, Language and Hearing Research, 42*, 1195-1204.

Cole, K.N., Dale, P.S. & Mills, P.E. (1992). Stability of the intelligence quotient-language quotient relation: Is discrepancy modelling based on a myth? American Journal on Mental Retardation, 97 (2) 131-143.

Cole, K.N., Schwartz, I.S., Notari, A.R., Dale, P.S., & Mills, P.E. (1995). Examination of the stability of two methods of defining specific language impairment. Applied Psycholinguistics, 16 (1) 103-123.

Conti-Ramsden, G., Botting, N., Simkin, Z., & Knox, E. (2001). Follow-up of children attending infant language units: outcomes at 11 years of age. *International Journal of Language and Communication Disorders, 36*(2), 207-219.

Dockrell, J.E. (2001). Assessing language skills in pre-school children. Assessing language skills in preschool children. *Child Psychology and Psychiatry Review*, **6**, 74- 85

Dodd, B. J., & Bradford, A. (2000). A comparison of three therapy methods for children with different types of developmental phonological disorder. *International Journal of Language and Communication Disorders, 35,* 189-209.

Dunn, L. M., Dunn, L. M., Whetton, C., & Burley, J. (1997). *British Picture Vocabulary Scale* (Rev. Ed.). Windsor, England: NFER-Nelson.

Ebbels, S, van der Lely, H & Dockrell, J.E. (in press). Intervention for verb argument structure in children with persistent SLI: a randomized control trial. *Journal of Speech Language and Hearing Research.*

Elliot, C. D., Smith, P., & McCulloch, K. (1997). *British Ability Scales Second Edition (BAS II).* Windsor: NFER-Nelson.

Elliot, C. D., Smith, P., McCulloch, K. (1996). *British Ability Scales Second Edition (BAS II)* Slough: NFER-Nelson.

Girolametto, L., Pearce, P. S., & Weitzman, E. (1996a). Interactive focused stimulation for promoting vocabulary in young children with delays: a Pilot Study. *Journal of Child Communication Development,* 17, 39-49.

Glogowska, M., Roulstone, S., Enderby, P., & Peters, T. J. (2000). Randomised controlled trial of community based speech and language therapy in preschool children. *British Medical Journal, 321*, 923-926.

Hake, R. R. (1998). Interactive engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*, 64-74.

Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society, A, 157,* 317 –356.

Haynes, C. & Naidoo, S. (1991). *Children with Specific Speech and Language Impairment. Clinics in Developmental Medicine 119*. Cambridge: MacKeith Press.

Howlin, P., & Cross, P. (1994).The variability of language test scores in 3- and 4-year-old children of normal non-verbal intelligence: a brief research report. *European Journal of Disorders of Communication, 29,* 279-288.

Irwig, L., Glasziou. P. & March, L. (1995) Ethics of n-of-1 trials. *Lancet*, 345. 369

Johnston, J. (2005). Re: Law, Garrett &Nye (2004a). "The efficacy of treatment for children with developmental Speech and language delay/disorder: A meta-analysis. *Journal of Speech. Language and Hearing Research 48,* 1114-1117.

Justice, L. (2006). Evidence-based practice, response to intervention and the prevention of reading difficulties. *Lang Speech Hear Serv Sch*, 37, 284-295.

Law J, Garrett Z, Nye C. (2003). Speech and language therapy interventions for children with primary speech and language delay or disorder. *The Cochrane Database of Systematic Reviews*, 3. (Art. No.: CD004110. DOI: 10.1002/14651858.CD004110).

Law, J. & Conti-Ramsden, G. (2000). Treating children with speech and impairments: six hours of therapy is not enough. *British Medical Journal*, 321, 908-909.

Law, J., Boyle, J., Harris, F., Harkness, A., & Nye, C. (1998). Child Health Surveillance: Screening for Speech and Language Delay. *Health Technology Assessment, 2*(9), 1-184.

Law, J., Boyle, J., Harris, F., Harkness, A., & Nye, C. (2000). Prevalence and natural history of primary speech and language delay: findings from a recent systematic review of the literature. *International Journal of Language and Communication Disorders*, *35*, 165-188.

Law, J., Lindsay, G., Peacey, N., Gascoigne, M., Soloff, N., Radford, J., & Band, S. (2001). Facilitating Communication between Education and Health Services: the provision for children with speech and language needs. *British Journal of Special Education.* 28, 3, 133-138.

Law, J., Dockrell, J., Williams, K., & Seeff, B. (2001). *The I CAN Early Years Evaluation Project.* http:qmuc.ac.uk/cihr

Law, J., Dockrell, J., Williams, K., & Seeff, B. (2004). Comparing specialist Early Years provision for speech and language impaired children with mainstream nursery provision in the UK– An application of the Early Childhood Environment Rating Scale (ECERS). *Child: Care, Health and Development*, *30*(2), 177-184.

Law, J. Dockrell, J.E, Castelnuovo, E., Williams, K, Seeff, B. & Normand, C.

(2006). Early Years Centres services for pre-school children with primary language difficulties: what do they cost, and are they cost-effective? *International Journal of Language and Communication Disorders*. **41,** 67-81.

Law, J., Tomblin, J.B. & Zhang, X. (in preparation). Characterising the growth trajectories of language impaired children between seven and eleven years

Leonard, L. (1998). *Children with specific language impairment.* Cambridge: MIT Press.

Matheny, N., & Panagos, J. M. (1978). Comparing the effects of articulation and syntax programmes on syntax and articulation improvement. *Language Speech and Hearing Services in Schools*, 9, 57-61.

McCauley, R., & Swisher, L. (1984). Psychometric review of language and articulation tests for children. *Journal of Speech and Hearing Disorders, 49*, 34-42.

McFadden, T. U. (1996). Creating language impairments in typically achieving children: The pitfalls of "normal" normative sampling. *Language Speech and Hearing Services in Schools, 27*(1) 3-9.

Nation, K., & Snowling, M. (1997). Assessing Reading Difficulties: The Validity and Utility of Current Measures of Reading Skill. *British Journal of Educational Psychology, 67*(3), 359-370.

Nye, C., Foster, S. H., & Seaman, D. (1987). Effectiveness of language intervention with the language/learning disabled. *Journal of Speech and Hearing Disorders, 52*, 348-357.

Petticrew, M. & Roberts, H. (2006). *Systematic reviews in the social sciences: a practical guide*. Oxford: Blackwell,

Pressley, M., & Harris, K. (1994). Increasing the quality of educational intervention research. *Educational Psychology Review, 6*(3), 191-208.

Pressley, M., Graham, S., & Harris, K.R (2006). The state of educational intervention research. *British Journal of Educational Psychology, 76,* 1-19.

Prieler, J. (2000) .Problems with Classical Change Scores (such as Measures of Learning Potential) and Their Resolution. *European Congress of Psychology*, London, England.

Renfrew, C. (1978). Renfrew Language Scales. Bicester, England: Speech Mark Ltd.

Reynell, J. & Huntley, M (1985).The Reynell Developmental Language Scales. NFER: Nelson: Windsor UK.

Robertson, S. B., & Weismer, S. E. (1999). Effects of treatment on linguistic and social skills in toddlers with delayed language development. *Journal of Speech, Language, & Hearing Research*, 42, 1234-1248.

Stothard, S. E., Snowling, M. J., Bishop, D. V. M., Chipchase, B. B., & Kaplan, C. A. (1998). Language impaired preschoolers: A follow-up into adolescence. *Journal of Speech, Language and Hearing Disorders*, *41*, 407-418.

Stuart, M. & Stainthorp, R. (2004). Viewpoint No.16. The assessment of reading: A theoretically motivated review of currently available tests. London: Institute of Education.

Wainer, H. (1991).Adjusting for differential base-rates: Lord's paradox again. *Psychological Bulletin, 109,*147-151.

Wertz, R. T. (2002). Evidence-based practice guidelines: not all evidence is created equal. *Journal of Medical Speech-Language Pathology, 10*(3), xi-xv.

Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology, 73*, 123-136.

Wright, D. B. (2006). Comparing groups in a before-after design: When t-test and ANCOVA produce different results. *British Journal of Educational Psychology, 76, 663-675*.

Zhang X. Y., & Tomblin, J. B. (2003). Explaining and controlling regression to the mean in longitudinal research designs. *Journal of Speech, Language, and Hearing Research, 46*(6), 1340-1351.

Zimmerman, I., Steiner, V. G., & Pond, R. E. (1992) .*Preschool Language Scale-3, UK Edition (PLS-3 (UK)).* Sidcup; The Psychological Corporation. UK adaptation: Boucher, J., & Lewis, V. (1997).

Zimmerman, I., Steiner, V. G., & Pond, R. E. (2002). Preschool Language Scale, Fourth Edition (PLS-4) English Edition. Sidcup; The Psychological Corporation.

TABLE 1

Language and Cognitive skills at entry to the provision for the two samples in Z-scores

	British Abilities Scale			Preschool Language Scale -3	
	Early Number concepts	Verbal ability	Nonverbal ability	Auditory	Expressive
Intervention Group Mean	-1.4	-1.5	-1.1	-1.6	-1.55
SD	1.0	.9	.8	.85	.55
Comparison Group Mean	-1.4	-1.3	-1.00	-1.35	-1.55
SD	1.1	1.1	.9	1.85	.9
Differences Between groups	t(89) =-2.46 *ns*	t(89) = -.984 *ns*	t(89)=-.848 *ns*	t (89)=-1.189 *ns*	t(89)=.083 *ns*

TABLE 2: Relationship between standard scores on language measures

Measure[3]	Verbal comprehension	Naming vocabulary	Verbal ability	PLS auditory
Naming vocabulary	.53			
Verbal ability	.89	.86		
PLS auditory	.70	.62	.76	
PLS expressive	.53	.58	.63	.56

[3] All correlations significant at the .001 level

TABLE 3

Cognitive and language subtests T scores from the British Abilities Scales at baseline and follow-up for the EYCs and Comparison group

BAS scales	Group	Time 1		Time 2		Mean gain
		Mean (SD)	95% confidence interval	Mean (SD)	95% confidence interval	
Block building	Intervention	40.45 (11.05)	37.54-43.35	42.69 (11.31)	39.72-45.66	2.24
	Comparison	41.36 (11.53)	37.26-45.47	43.55 (10.79)	39.72-47.37	2.18
Picture similarities	Intervention	38 (8.56)	35.75-40.25	41.79 (6.89)	39.98-43.61	3.79
	Comparison	41.85 (13.56)	37.04-46.66	40.52 (11.05)	36.60-44.43	-1.33
Verbal comprehension	Intervention	33.45 (10.53)	30.68-36.22	39.52 (8.59)	37.26-41.78	6.07
	Comparison	36.58 (11.22)	32.60-40.55	38.15 (9.05)	34.94-41.36	1.58
Naming vocabulary	Intervention	37.17 (8.95)	34.82-39.43	39.41 (8.26)	37.24-41.58	2.24
	Comparison	37.94 (11.63)	33.82-42.06	41.24 (9.41)	37.90-44.58	3.30

Number concepts	Intervention	36.16 (9.49)	33.66-38.65	39.79 (8.47)	37.57-42.02	3.63
	Comparison	35.64 (5.58)	32.24-39.03	39.24 (10.88)	35.38-43.10	3.60

TABLE 4

Preschool Language Scale Standard Scores at baseline and follow-up for the EYC and Comparison group

PLS	Group	Time 1		Time 2		Gain
		Mean (range)	95% confidence interval	Mean (range)	95% confidence interval	
Auditory	Intervention	76.36 (12.87)	72.98-79.75	75.74 (14.18)	72.02-79.47	-.6207
	Comparison	80.00 (15.89)	74.37-85.63	81.24 (17.95)	74.88-87.61	1.24
Expressive	Intervention	77.31 (7.85)	75.25-79.37	76.26 (10.22)	73.57-78.95	-1.05
	Comparison	77.12 (13.87)	72.20-82.04	80.64 (10.97)	76.75-84.53	3.52

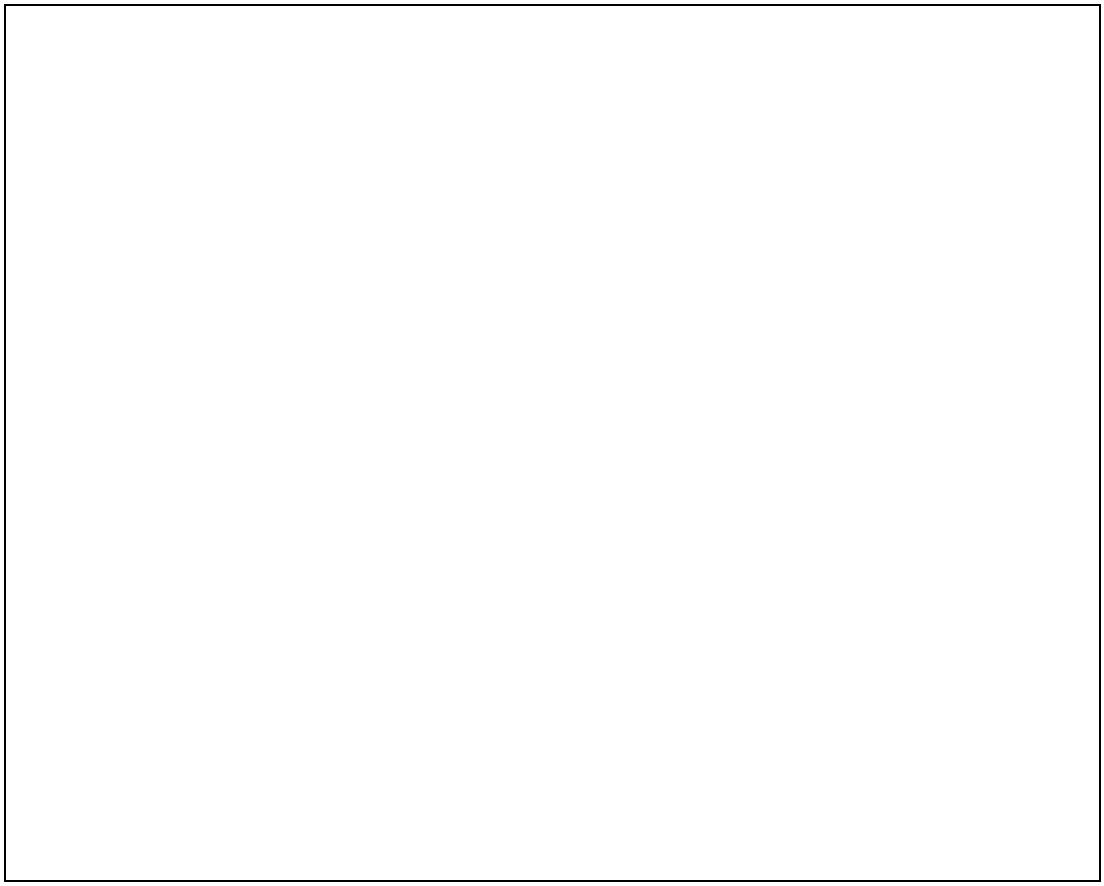

Figure 1 Relationship between gain in verbal comprehension and baseline comprehension score for EYC and Comparison group.