doi: 10.23965/AJEC.42.1.12

Establishing agreement between parent-reported and directly-measured behaviours

Shannon K. Bennetts
The University of Melbourne
Murdoch Childrens Research Institute
La Trobe University

Fiona K. Mensah
The University of Melbourne
Murdoch Childrens Research Institute
The Royal Children's Hospital, Melbourne

Elizabeth M. Westrupp
La Trobe University
The University of Melbourne
Murdoch Childrens Research Institute

Naomi J. Hackworth
La Trobe University
Murdoch Childrens Research Institute

Jan M. Nicholson
La Trobe University
Murdoch Childrens Research Institute
Queensland University of Technology

Sheena Reilly
Griffith University
Murdoch Childrens Research Institute
The University of Melbourne

THE QUALITY AND ACCURACY OF research findings relies on the use of appropriate and sensitive research methods. To date, few studies have directly compared quantitative measurement methods in the early childhood field and the extent to which parent-reported and directly-measured behaviours agree is unclear. Existing studies are hampered by small sample sizes and the use of statistical techniques which quantify the magnitude of association between measures (e.g. correlations), but not agreement. Here we review the limitations of existing method comparisons and suggest how alternative statistical approaches such as the Bland-Altman Method and ordinary least products regression can be readily applied in the early childhood context. Understanding agreement (and disagreement) between measurement methods has potential to reduce research costs and improve data quality, with important implications for researchers, clinicians and policy-makers.

Introduction

Quantitative measurement methods in the early childhood field commonly include parent-report (e.g. parent reporting on child's ability), direct observation and direct assessment. While parent-reported measures are relatively inexpensive, they are vulnerable to parental subjectivity (Law & Roy, 2008). Direct measures, such as observed parent and child behaviours or standardised assessments, are often seen as the gold standard, allowing more objective measurement (Hawes & Dadds, 2006), but are costly and time-consuming and may not generalise beyond the clinical assessment room. Method comparisons have therefore sought to quantify the agreement between these methods, with the potential for significant cost and time-savings if methods are demonstrated to be interchangeable. Current evidence is limited by small sample sizes and the inappropriate use of correlations (Bland & Altman, 1986). This paper reviews statistical techniques used in existing early childhood method comparisons, focusing on child language and parent–child interactions, and discusses alternative statistical approaches.

In the early childhood field, measurement may involve parent-report, direct assessments and observations, or qualitative methodologies such as parent interviews. Where the selection of quantitative measures is limited by time, cost or resources, researchers may attempt to model or understand 'method effects'—the systematic variability associated with particular measurement methods (Fiske, 1987). While there is inevitable discrepancy between any two methods (Hawes & Dadds, 2006), understanding the level of agreement or disagreement is crucial for high quality and efficient measurement. Such information can inform decisions to administer one method rather than two, or the selection of a measure that is associated with less expense or lower research burden.

Method comparisons typically have one of two aims: (1) detecting bias; and (2) calibrating one method against another (Bland & Altman, 1986; Ludbrook, 2010a). Calibration is used when a new method is compared to a *known* ('true') quantity such as height or weight (Bland & Altman, 1986). In early childhood research, constructs being measured typically have a high degree of subjectivity, for example, children's expressive language abilities, or the degree of warmth parents show towards their child. Although we endeavour to measure such constructs objectively using scales (e.g. Clinical Evaluation of Language Fundamentals), no single measure is likely to provide an unequivocally 'correct' measurement. Therefore, method comparison studies aim to detect bias, to quantify the difference between methods. This allows researchers or clinicians to determine whether this difference permits interchangeability of methods in a given context (Ludbrook, 2002). A *lack* of agreement may indicate that two methods may be measuring different constructs, or that one or both methods contain substantial error (Cox, 2006).

The purpose of this paper is three-fold to: (1) review existing methods for the comparison of parent-reported and direct measurement methods; (2) define agreement as it pertains to method comparisons; and (3) present alternative and more robust techniques for quantifying agreement between methods in early childhood. By way of example, we focus on continuous measures of child language and parent–child interactions; categorical variables are beyond the scope of this paper.

Measurement methods in early childhood research

Parent-reported and direct measures are commonly used to measure parent and child behaviour. The former are relatively low-cost, allowing the data to be collected rapidly (Law & Roy, 2008). However, this method may produce data which is influenced by factors such as parent mental health, personality, expectations, socioeconomic status or social desirability (Feldman et al., 2000; Gilmore & Cuskelly, 2011; Hayden, Durbin, Klein & Olino, 2010; Zaslow et al., 2006). Direct measures such as observations or standardised assessments are often considered to be the 'gold standard', allowing researchers to observe behaviours in real time and reducing the impact of parental biases (Aspland & Gardner, 2003). Direct measures can also be time-consuming and costly to administer, participants may modify their behaviour in the presence of the observer and findings may not generalise well to other settings (Gardner, 2000). Given these limitations, researchers frequently collect data using multiple methods (Zaslow et al., 2006). Particularly within low-income populations, there is a tendency for weaker associations between parent-reported and directly-measured behaviours compared to within higher-income samples (Pan, Rowe, Spier & Tamis-LeMonda, 2004). This may be due to factors such as maternal education, maternal age and ethnicity, which can influence the ways that parents interpret and respond on parent-reported measures. While the use of multiple methods can provide data which together explain a greater amount of variance in child outcomes than a single method alone (Zaslow et al., 2006), this is not always feasible given practical and financial constraints and is burdensome for participants.

Comparing measurement methods

In the early childhood field, method comparisons have primarily focused on comparing parent-reported measures with directly-observed or directly-assessed measures of child temperament (Hayden et al., 2010; Olino, Durbin, Klein, Hayden & Dyson, 2013; Seifer, Sameroff, Barrett & Krafchuk, 1994), parenting behaviours such as responsiveness and control (Arney, 2004) or praise and criticism (Hawes & Dadds, 2006) and child language (Ring & Fenson, 2000; Sachse & Von Suchodoletz, 2008). Measurement methods are typically compared using correlations. We provide a brief review of method comparisons below for the measurement of child language and parent–child interaction. Evidence suggests that parental behaviours during early childhood, such as responsiveness, are predictive of children's later language outcomes (Hudson, Levickis, Down, Nicholls & Wake, 2015).

Child language

Associations between parent-reported and directly-assessed measures of child language are generally moderate to strong. For example, in an American study, Ring and Fenson (2000) compared parent-reported data from the commonly used MacArthur Bates Communicative Development Inventory (CDI) with a direct laboratory-based child assessment of comprehension and production (n = 40). Correlations between parent-reported and direct assessments were moderate and were higher for production (r = 0.67) than comprehension (r = 0.55). In a German study, scores on the German version of the CDI (ELFRA-2) were compared with scores on a direct laboratory-based child assessment (SETK-2) in a sample of typically developing (n = 47) and slow-to-talk two-year-olds (n = 70) (Sachse & Von Suchodoletz, 2008). Again, stronger rank order correlations were found between measures of word production (r = 0.87) than between measures of word comprehension (r = 0.40). These associations are consistent with findings that indicate parents are more able to accurately report on what their child says but have more difficulty reporting what their child understands (Eriksson, Westerlund & Berglund, 2002).

Parent–child interaction

Associations between parent-reported and directly-measured parent–child interaction appear to be generally weaker than associations for child language. For example, Arney (2004) administered the 30-item Child Rearing Practices Questionnaire and the 30-item Parenting Scale. Parallel constructs were then compared using observational data coded from video-taped, semi-structured play activities with parents and children in the home (*n* = 68). A modest correlation was found between parent-reported and observed parental warmth (*r* = 0.36), with very weak or negligible correlations for parenting inconsistency (*r* = 0.01), permissiveness (*r* = 0.05), over-reactivity (*r* = 0.11), reasoning (*r* = 0.12) and child obedience (*r* = 0.04). Such findings highlight the difficulties in accurately capturing parent and child behaviours. Rare behaviours might be most appropriately captured using parent-report methods. However, measuring socially undesirable behaviours such as over-reactivity can be problematic using both methods; these behaviours may be observed infrequently in an observational setting and are also susceptible to under-reporting by parents.

Limitations of existing research

Method comparison studies in the early childhood field are scant. Currently, there is insufficient evidence to permit sound conclusions to be drawn about agreement between parent-reported and directly-measured behaviours, and existing correlational data suggests substantial differences between what are purported to be measures of the same construct. Review of the existing literature reveals two primary limitations: small sample sizes and reliance on correlations.

Sample size

Sample sizes of method comparisons have generally been small (around *n* = 50–70), limiting a study's precision to accurately detect the magnitude of bias. This is not unexpected, given that direct measures of behaviour require significant investment of time and financial resources. Bland (2004) recommends that at least *n* = 100 is a reasonable sample size for method comparisons, therefore many existing studies may have insufficient sample sizes to accurately quantify agreement between methods. One notable exception is Olino and colleagues (2013) who drew on three cohort studies to produce a substantially larger sample for their investigation of child temperament (*n* = 865).

Use of correlations versus assessing agreement

The Pearson's Correlation Coefficient (*r*) is commonly used to compare parent-reported and direct measures. This is problematic because Pearson's Correlation provides information about the degree of *association* between variables, but not on their level of *agreement* (Altman & Bland, 1983). The Pearson's correlation provides a single figure to quantify the degree of linear association between two continuous variables (i.e. how closely a plot of the variables would follow a straight line) (Kirkwood & Sterne, 2013). However, it is possible for two measures to be linearly related *and* demonstrate poor agreement (Eadie et al., 2014). Agreement is a measure of the 'closeness' between two readings (Barnhart, Haber & Lin, 2007). This 'closeness' can vary across the distribution of scores, a characteristic which cannot be described using a correlation. Quantifying this agreement accurately is an essential step in determining whether two methods could be used interchangeably.

In the early childhood field, correlations continue to be widely used for method comparisons contrary to the advice of statisticians (e.g. Bland & Altman, 1990; Carstensen, Simpson & Gurrin, 2008; Cox, 2006; Ludbrook, 2002). Ludbrook (2002) states that there is 'universal agreement … [amongst biostatisticians] that the Pearson product–moment correlation coefficient (*r*) is valueless as a test for bias' (p. 527). Proliferation of such techniques in method comparisons may contribute to inaccurate perceptions of agreement.

To illustrate how correlations can be misleading, consider three methods designed to measure the same construct. Methods A, B and C are administered to 20 individuals (see Table 1). We can see from Table 1 that Method B consistently produces scores 30 units higher than Method A. This will result in a high correlation. Taken in isolation,we may be tempted to conclude that this correlation suggests the methods are in 'agreement'; however, this is not the case. Depending on the context, a difference of 30 units may be quite substantial and indicative of considerable discrepancy between methods. Importantly, it should be noted that the level of agreement deemed satisfactory should be governed by the context in which measurement is occurring (Kirkwood & Sterne, 2013) and resulting implications.

For some purposes (e.g. making decisions about individual children who should receive education assistance programs), very high agreement may be necessary before we would consider using two methods interchangeably. In other contexts (e.g. monitoring average developmental change in a population over time), poorer agreement may be acceptable. Correlations also provide an estimation of association for a total study population, which prevents the identification of varying levels of bias across the distribution of scores (Hopkins, 2004). For example, there may be different levels of agreement between methods for children with below average, average, or above average expressive language abilities. This level of detail cannot be determined with correlations but is essential for understanding what measures may be appropriately administered and to whom.

Table 1. Example of participant scores across three methods (*n* = 20)

Method A	Method B	Method C
29	59	37.7
36	66	46.8
38	68	49.4
42	72	54.6
47	77	61.1
49	79	63.7
51	81	66.3
54	84	70.2
56	86	72.8
57	87	74.1
61	91	79.3
64	94	83.2
65	95	84.5
65	95	84.5
73	103	94.9
78	108	101.4
82	112	106.6
84	114	109.2
89	119	115.7
94	124	122.2

Quantifying bias

Differences between methods can be attributed to two main sources: (1) measurement error; and (2) bias. Measurement error is 'random' and occurs due to unpredictable factors (e.g. a parent or child is feeling tired during the assessment, or natural variability in a child's performance). Bias occurs when there are systematic differences between methods across the population of interest (e.g. parents consistently rating their child's language comprehension higher than scores on a direct assessment) (Kirkwood & Sterne, 2013). Method comparison studies focus on quantifying the *differences* between methods; this magnitude of bias is crucial for evaluating interchangeability (Ludbrook, 2002). Bias can occur in two forms: 'fixed' bias produces higher or lower scores by a *constant* amount; and 'proportional bias' produces scores which vary *proportionally* in relation to the level of the reference variable (Ludbrook, 1997).

Statistical techniques vary in their ability to detect fixed and proportional bias. Consider the example of Method B consistently producing scores 30 units higher than Method A (Table 1). This is an example of fixed bias. In this situation, it may be possible to calibrate the measures by adding an additional 30 units to Method B scores.

In contrast, proportional bias occurs when the magnitude of bias varies across the distribution of scores, and as such, a simple calibration is not feasible. These phenomena are illustrated in Figures 1a and 1b. 'Perfect' agreement is denoted by the dotted line. Figure 1a represents the example of fixed bias described above. In this example, a consistent difference of 30 units between Method A and B produces a perfect correlation (*r* = 1.00), even though such a discrepancy is likely to indicate poor agreement between methods. Figure 1b illustrates proportional bias, whereby the magnitude of bias between Methods A and C increases proportionally as scores increase. The correlation coefficient represents the association for the total study population, but does not account for the variation across the distribution of measures (i.e. less bias at the lower end and greater bias at the upper end). In both examples, a perfect correlation is produced, but this alone provides no indication of systematic bias or the type of bias present.

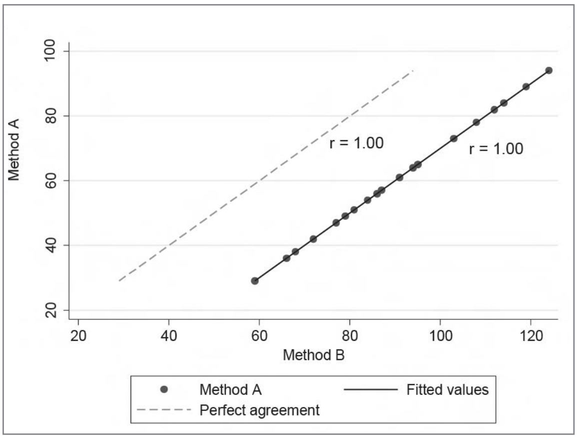


Figure 1a. Example of fixed bias

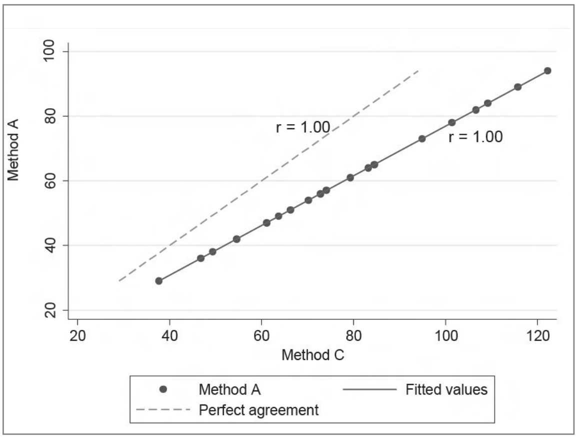


Figure 1b. Example of proportional bias

Alternative approaches to identify bias and determine interchangeability

Table 2 summarises the most common statistical techniques for method comparisons, their primary purposes and suitability for the detection of bias. Terminology varies across disciplines, so alternative terms for the same technique are provided.

Alternative correlation coefficients

There remains a lack of consensus about the usefulness of alternative correlation coefficients such as the Intraclass Correlation Coefficient (ICC, r_1) and the Concordance Correlation Coefficient (CCC, *pc*), as described in Table 2. Bland and Altman (1990) argue against using any correlation, suggesting that while useful for validity studies (such as determining construct validity of a new measure), a single correlation coefficient cannot be used to determine interchangeability between methods. Lee, Koh and Ong (1989) suggested that if the lower limit of the 95 per cent confidence interval of the ICC was at least 0.75, this would provide evidence of potential interchangeability between methods. The ICC was purported to measure individual participant agreement, with a view to determining potential interchangeability. This suggestion was met with criticism by Bland and Altman (1990, p. 338), who noted that the ICC requires several assumptions to be met, such as equal measurement error of both methods; an assumption they deemed 'unjustified'. In 1989, Lin proposed another correlation coefficient for method comparisons: the CCC (*p*c, see Table 2) which is said to account for both systematic bias and random error (Lin, 1989). The CCC received criticism for its inability to determine whether poor agreement is due to homogeneity of the sample, systematic bias between methods, or random error between methods (Atkinson & Nevill, 1997).

Bland-Altman Method

Given the common conceptual misunderstandings of 'agreement' and resultant applications of inappropriate statistical techniques, two primary approaches have emerged; (1) visual techniques which involve mean-difference plots; and (2) linear regression techniques. Over the past three decades, medical statisticians Martin Bland and Douglas Altman have advocated for the use of a mean-difference plot, now commonly termed the 'Bland-Altman Method' or the 'Limits of Agreement Approach' (see Table 2). Following their original paper in *The Statistician* (Altman & Bland, 1983), their 1986 paper in *The Lancet* saw this method rise to prominence; the latter is one of the most highly cited statistical papers, with over 35 000 citations to date (Google Scholar Search, http://scholar.google.com.au, 2016). Bland and Altman (1990, p. 339) suggest that an exploratory and graphical approach is most appropriate in determining agreement, arguing that although temptingly convenient, it is 'impossible to summarise agreement adequately using a single number'. Based on similar work by Eksborg (1981), the Bland-Altman Method aims to quantify the average difference between two methods. It allows us to examine individual variability in those differences according to the estimated 'true' value, and to determine whether the limits of agreement between two methods are narrow enough to permit interchangeability (Carstensen, 2010). In other words, are the measures similar enough to be used interchangeably?

While there is no consensus on the 'ideal' statistical method for quantifying bias in method comparisons (Mullineaux et al., 1999), the Bland-Altman Method is appealing for its simplicity and visual nature. This technique involves calculating the mean of the two methods for each individual, as well as the difference between methods for each individual. These mean and difference scores are then plotted against each other (Figure 2 on page 111). A line of mean difference represents the average difference between the two methods; the estimated bias. 'Perfect' average agreement is indicated by a mean line of zero, together with data points positioned along this line. Upper and lower limits, termed the 'limits of agreement', are calculated as the mean difference, plus or minus twice the standard deviation of the differences. Approximately 95 per cent of the data will fall within this range if the differences are normally distributed. Wider limits of agreement indicate poorer agreement and narrower limits of agreement indicate stronger agreement. Unlike correlations, the Bland-Altman Method enables the identification of systematic bias; we can detect differences in agreement along the distribution of scores. For example, it's possible to ask; is there poorer agreement between Methods A and B for children with poorer language compared to those with typical or advanced language? Or perhaps vice versa? Alternatively, is agreement similar for all children?

Table 2. Statistical techniques used to compare measurement methods

Method	Alternative terms	Purpose	Suitability for detecting bias to evaluate interchangeability	Suggested references
Pearson's Correlation Coefficient	*r*; Pearson's *r*; Product Moment Correlation Coefficient; Interclass Correlation	Assesses strength of linear association.	Useful exploratory analysis but does not permit the detection of bias.	Altman and Bland (1983); Bland and Altman (1986)

Continued on next page ...

... continued from previous page.

Method	Alternative terms	Purpose	Suitability for detecting bias to evaluate interchangeability	Suggested references
Intraclass Correlation Coefficient	*ICC; r*1	Assesses strength of association between variables within classes.	Useful exploratory analysis but does not permit the detection of bias.	Bland and Altman (1990, 1999); Lin (1989); Nickerson (1997)
Spearman's Rank Correlation Coefficient	Spearman's Rho; Grade Correlation, rsp;	Assesses the strength of association using the Pearson's Correlation between ranked variables. Useful for non-linear data.	Not appropriate. Does not permit the detection of bias.	Kirkwood and Sterne (2013)
Concordance Correlation Coefficient	Lin's Coefficient; Rho c; c; CCC	Assesses the agreement on a continuous measure by two persons or methods, correcting for systematic bias and random error.	Useful exploratory analysis, but arguably unsuitable as a single indicator of agreement.	Lin (1989); McBride (2005)
T-test	Student's T-test	Determines if the means of two methods are significantly different.	Useful exploratory analysis to assess bias between means but does not permit the assessment of agreement across the distribution of scores.	Kirkwood and Sterne (2013)
Bland-Altman Method	Mean-Difference Plot; Tukey Mean-Difference Plot; BA Method; Limits of Agreement Approach	Quantifies bias between methods by plotting the mean of two methods against the difference between the methods.	Appropriate. Provides a visual means of examining agreement across the distribution of scores; however it may not distinguish between fixed and proportional bias.	Bland and Altman (1986); Carstensen et al. (2008); Eadie et al. (2014)
Ordinary Least Squares Regression	Simple Linear regression; OLS	Estimates the best-fitting line to describe an association by plotting a regression line which minimises the squared sum of the vertical residuals.	Not usually appropriate. Can be used if values for one method are fixed by the study design.	Mullineaux, Barnes and Batterham (1999)
Ordinary Least Products Regression	OLP; Reduced Major Axis (RMA) Regression; Standardised Major Axis Regression; Standardised Principal Component Regression; Geometric Mean Regression	Estimates the best-fitting line to describe an association by plotting a regression line which minimises the sum of the products of the vertical and horizontal residuals.	Appropriate. Can be used to distinguish between fixed and proportional bias. Unlike OLS, values for both methods are free to vary and may be subject to error.	Ludbrook (1997, 2010b); Mullineaux et al. (1999)
Weighted Least Squares Regression	WLS	Estimates the best-fitting line to describe an association by plotting a regression line which minimises the weighted sum of squares.	Appropriate if heteroscedasticity is present (scatter of points around the regression line is inconsistent across the range of values). Can be used if values for one method are fixed by the study design and are not subject to error.	Ludbrook (2010b)
Weighted Least Products Regression	WLP; Weighted Major Axis Regression (WMA); Weighted LPR	Estimates the best-fitting line to describe an association by plotting a regression line which minimises the weighted sum of the products of the vertical and horizontal residuals.	Appropriate if heteroscedasticity is present. Similar to OLP, in that values for both methods are free to vary, however WLP can be used if the heteroscedasticity assumption is violated.	Ludbrook (2010b)

Traditionally, the Bland-Altman Method has been used to compare alternative medical instruments, such as methods to measure blood pressure (Bland & Altman, 1986). It can also be readily applied to other comparisons, such as parent–child or parent–teacher responses on the same scale, alternative measures of subjective constructs such as health-related quality of life, or stability and participant variability in outcomes between time points (e.g. Eadie et al., 2014; Gabbe et al., 2010; Quercioli et al., 2009). In the early childhood field, measures are commonly reported on different scales, preventing a direct and meaningful comparison. In this case, scores on each method can firstly be converted to z-scores. This results in a plot where the mean difference is around zero. The 95 per cent limits of agreement can then be interpreted in terms of standard deviations.

It should be emphasised that there is no criterion for determining 'good' or 'poor' agreement. Bland-Altman plots and limits of agreement should be interpreted in the context in which measurement has taken place, with consideration to purpose and implications of measurement. For example, we may want stronger agreement when administering clinical assessments with an individual child, but poorer agreement may be acceptable at a population level. If there is unacceptably poor agreement between two measures, we may opt for the 'gold standard' measure, such as the Clinical Evaluation of Language Fundamentals (CELF), instead of a parent-reported measure. Conversely, if strong agreement is obtained, the less onerous measure can be administered, such as a brief parent-reported measure. In this way, method comparisons can guide the allocation of limited funds and resources, minimising the need to administer multiple time-consuming and expensive assessments.

To our knowledge, the only application of the Bland-Altman method in the early childhood field is that of Eadie and colleagues (2014) who compared scores from the same direct child language measure (CELF) administered to children at age four and five. Although this is not an example of a parent-reported and direct measure comparison, it demonstrates the flexibility of the Bland-Altman Method for a range of comparative purposes. Figure 3 shows a Bland-Altman plot from Eadie et al., 2014. The data shows little evidence of bias; the points are dispersed fairly evenly across the distribution of Expressive Language Index Scores and the mean difference lies close to zero (mean difference: –1.25 standard score units [solid line], 95% CI: –1.86, –0.64). This suggests that the scores were, on average, slightly lower at five years compared to four years.

Visual examination of plots can also reveal any patterns of bias across the range of values. Two common effects are demonstrated using conceptual schemas in Figures 4a and 4b. For example, a funnelling effect (Figure 4a) suggests that agreement is stronger at one end and poorer at the opposite end. Likewise, a diamond-shaped distribution (Figure 4b) suggests that agreement is poorest in the middle of the distribution and strongest at the upper and lower extremes. Using the data from Table 1, Figure 5a shows the data points along a horizontal line, indicating that the difference between Methods A and B is consistently 30 units across the spectrum of scores. Figure 5b shows how Method C produces progressively higher scores than Method A. Such plots represent a relatively straightforward avenue for identifying variations in agreement between methods across the distribution of scores.

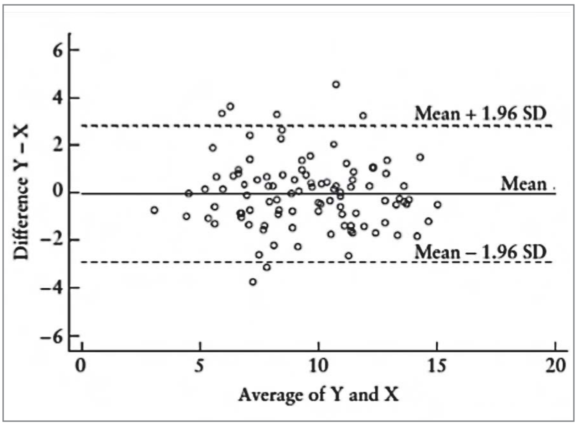


Figure 2. Bland-Altman plot for Methods X and Y, with mean difference and 95 per cent limits of agreement indicated (Bland & Altman, 2003, p. 90)

Reproduced with permission (Copyright © 2003 ISUOG. Published by John Wiley & Sons, Ltd).

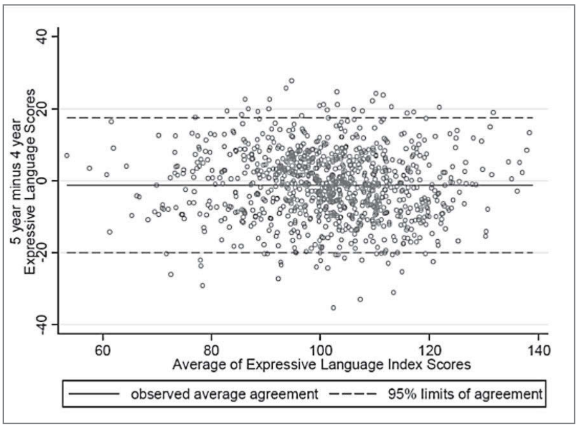
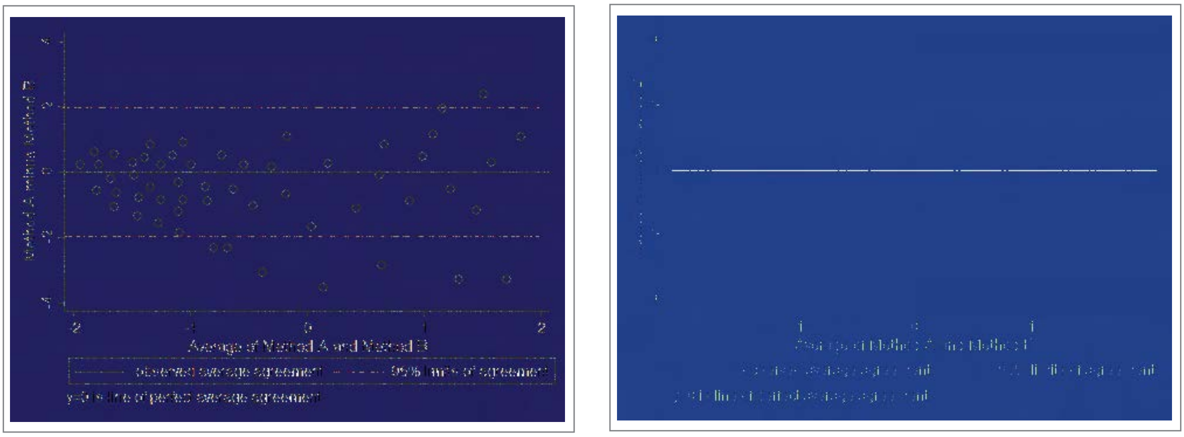


Figure 3. Bland-Altman plot of the difference against the average of four- and five-year Expressive Language Index scores (Eadie et al., 2014, p. 221)

Reproduced with permission (Copyright © 2013 Royal College of Speech and Language Therapists).



Figures 4a and 4b. Conceptual schemas showing systematic bias: funnel-shaped (left) and diamond-shaped (right) distributions

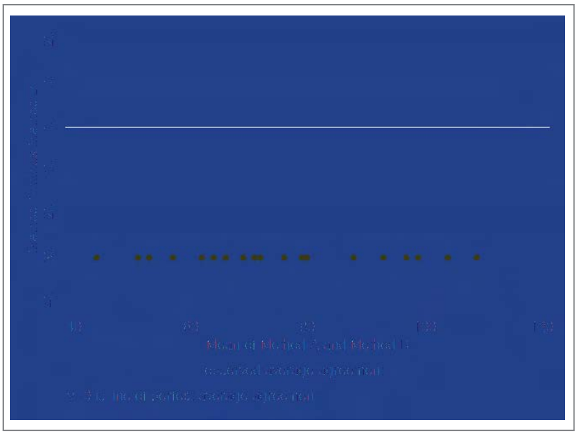


Figure 5a. Bland-Altman plot of Methods A and B (mean difference = –30.00; 95% limits of agreement: –30.00, –30.00)

Note that the limits of agreement are not visible because the upper and lower limits are exactly aligned.

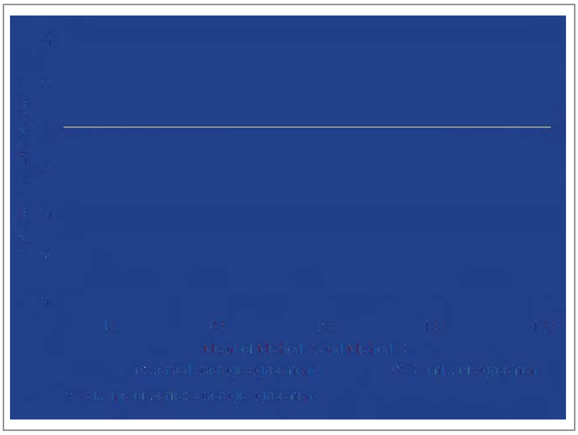


Figure 5b. Bland-Altman plot of Methods A and C (mean difference = –18.21; 95% limits of agreement: –28.99, –7.43)

Although Bland and Altman have been credited with reduced use of correlations and rapid uptake of mean-difference plots in the biomedical field (Ludbrook, 2010a), their approach remains uncommon in many non-medical fields, including early childhood. This may be due to a lack of exposure to interdisciplinary literature. Many researchers follow the methods published in their own field (Miles & Banyard, 2007), perpetuating the absence of mean-difference plots from use and subsequent publication.

Linear regression techniques

Other researchers have advocated for the use of linear regression techniques. John Ludbrook (e.g. 2002, 2010b) challenged the application of the Bland-Altman Method, citing its inability to differentiate between fixed and proportional bias and recommending regression as the preferred option. Regression modelling has a long history of use for method comparisons, particularly in clinical chemistry (Dewitte, Fierens, Stöckl & Thienpont, 2002), yet has been seldom used for this purpose in the early childhood field. The commonly used Ordinary Least Squares regression (OLS) is typically not appropriate for method comparisons in this field. This is because OLS assumes that one method is fixed (i.e. has no bias/error); an assumption that may not be met when subjective measures such as child language or parent–child interaction are concerned.

Perhaps the most suitable is Ordinary Least Products regression (OLP), which allows both measures to include some error. OLP applies a regression line that minimises the sum of the vertical and horizontal residuals from the line.

This is distinct from OLS, in which only the sum of the squared *vertical* residuals from the line is minimised. The primary strength of OLP is that, unlike the Bland-Altman Method, it allows researchers to distinguish between fixed and proportional bias (Kramer, Cumming, Churilov & Bernhardt, 2013). Fixed bias is estimated via the intercept of the regression line and proportional bias is indicated by the slope of the regression line (Mullineaux et al., 1999). This method is illustrated in Figures 6a and 6b, again using the illustrative data from Table 1. The regression coefficients and 95 per cent confidence intervals around these coefficients can then be used to draw conclusions about the presence of bias (Ludbrook, 2010a). If the regression coefficient for the intercept differs from zero and the 95 per cent confidence interval does not include zero, there is evidence of fixed bias (see Figure 6a). If the regression coefficient for the slope differs from one and the 95 per cent confidence interval does not include one, there is evidence of proportional bias (see Figure 6b).

In some cases there may be heteroscedasticity present, whereby the scatter of points around the regression line is not consistent across the range of values (Lumley, Diehr, Emerson & Chen, 2002). This characteristic violates the assumptions of OLS and OLP. If there is heteroscedasticity but the values of one method are fixed by the research design, a weighted version of OLS will suffice (i.e. Weighted Least Squares). This is unlikely to be the case in early childhood research, where there is typically no 'true' value and both methods are likely to contain some degree of measurement error and bias. Therefore, a weighted version of OLP (i.e. Weighted Least Products or WLP) can be used, while still allowing ease of interpretation. WLP involves dividing the *y* residuals by *x*, and dividing the *x* residuals by *y*, and minimising their products (see Ludbrook, 1997).

Contributing to confusion and a hesitancy to apply the abovementioned techniques is the varied terminology. As described in Table 2, OLP is also known as Reduced Major Axis (RMA) or Geometric Mean regression. For an extensive review of linear regression techniques for method comparisons, see Ludbrook (2010b).

Summary and recommendations

To date, efforts to compare parent-reported and directly measured behaviours in early childhood have been limited by inappropriate statistical methods, such as correlations. Continued misuse of statistical techniques will hinder efficient parent–child assessment and the generation of high-quality evidence. This is therefore an important issue for all professionals within the field, from researchers conducting large trials, to clinicians administering assessments with young children and their families, and policy-makers reviewing the scientific literature.

The Bland-Altman Method and OLP regression represent appropriate statistical techniques for determining agreement and potential interchangeability between measurement methods in the early childhood field. Evidence of strong agreement between two measures may warrant the use of only the parent-reported measure, whereas evidence of poor agreement may warrant use of only the accepted 'gold standard' direct measure. Scores on different scales can be converted to z-scores and interpreted in units of standard deviation. These techniques enable interpretation of differences in findings from multiple collection methods and the strategic selection of optimal measurement methods to support more sustainable and low-cost measurement. The Bland-Altman Method is relatively simple to execute and allows for agreement to be scrutinised visually, including

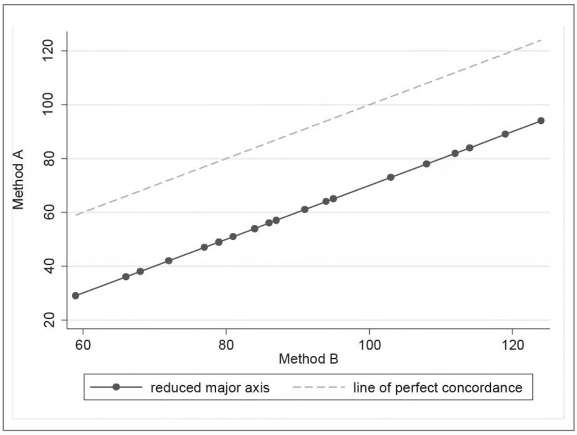


Figure 6a. OLP plot showing evidence of fixed bias (Slope = 1.00; Intercept = –30.00)

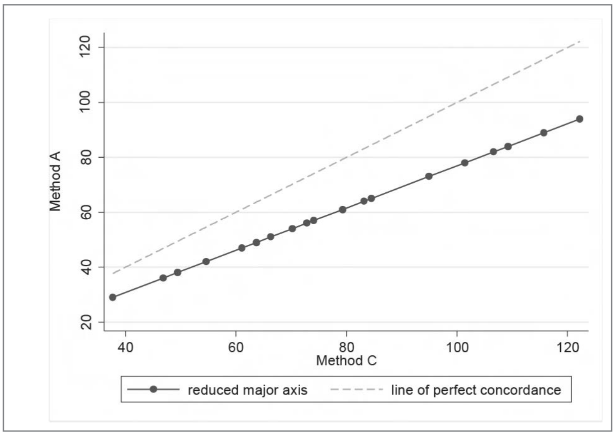


Figure 6b. OLP plot showing evidence of proportional bias (Slope = 0.77 ; Intercept = 0.00)

variability in agreement across the distribution of scores. OLP regression is appropriate when measurement error is expected for both methods, and unlike the Bland-Altman Method, can distinguish between fixed and proportional bias. The Concordance Correlation Coefficient, while arguably inadequate in isolation, may be used as part of a multi-method approach to determining agreement. We encourage journal reviewers to support the use of these techniques, rather than permitting continued use of the Pearson's Correlation Coefficient for the analysis of method comparisons.

Funding and acknowledgements

Ms Shannon Bennetts and Dr Westrupp were supported by the NHMRC-funded Centre of Research Excellence in Child Language (#1023493); Dr Fiona Mensah was supported by an NHMRC Early Career Fellowship (#1037449) and Career Development Fellowship (#1111160); Professor Sheena Reilly was supported by an NHMRC Practitioner Fellowship (#1041892) and Ms Bennetts, Dr Westrupp, Dr Hackworth and Professor Jan Nicholson were supported by the Australian Communities Foundation through the Roberta Holmes Transition to Contemporary Parenthood Program (Coronella sub-fund) at La Trobe University. The contents of the published material herein are the sole responsibility of the authors and do not reflect the views of the NHMRC. Research at the Murdoch Childrens Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program. The authors gratefully acknowledge Professor John Carlin, Associate Professor Patricia Eadie and Dr Cattram Nguyen for comments on this paper. The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Altman, D., & Bland, M. (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician), 32*(3), 307–317. doi: 10.2307/2987937

Arney, F. (2004). *A comparison of direct observation and self-report measures of parenting behaviour* (PhD disertation). Adelaide University, Adelaide. Retrieved from http://hdl.handle.net/2440/37713.

Aspland, H., & Gardner, F. (2003). Observational measures of parent–child interaction: An introductory review. *Child and Adolescent Mental Health, 8*(3), 136–143. doi: 10.1111/1475-3588.00061

Atkinson, G., & Nevill, A. (1997). Comment on the use of Concordance Correlation to assess the agreement between two variables. *Biometrics, 53*(2), 775–777.

Barnhart, H. X., Haber, M. J., & Lin, L. I. (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics, 17*(4), 529–569. doi: 10.1080/10543400701376480

Bland, M. (2004). *How can I decide the sample size for a study of agreement between two methods of measurement?* Retrieved 19 May, 2016, from www-users.york.ac.uk/~mb55/meas/sizemeth.htm.

Bland, M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307–310. doi: 10.1016/S0140-6736(86)90837-8

Bland, M., & Altman, D. (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers in Biology and Medicine, 20*(5), 337–340. doi: 10.1016/0010-4825(90)90013-F

Bland, M., & Altman, D. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135–160. doi: 10.1177/096228029900800204

Bland, M., & Altman, D. (2003). Applying the right statistics: Analyses of measurement studies. *Ultrasound in Obstetrics and Gynecology, 22*(1), 85–93. doi: 10.1002/uog.122

Carstensen, B. (2010). *Comparing Clinical Measurement Methods: A Practical Guide*. Chichester, UK: John Wiley & Sons Ltd.

Carstensen, B., Simpson, J., & Gurrin, L. (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics, 4*(1), Article 16.

Cox, N. J. (2006). Assessing agreement of measurements and predictions in geomorphology. *Geomorphology, 76*(3–4), 332–346. doi: 10.1016/j.geomorph.2005.12.001

Dewitte, K., Fierens, C., Stöckl, D., & Thienpont, L. M. (2002). Application of the Bland–Altman plot for interpretation of method-comparison studies: A critical investigation of its practice. *Clinical Chemistry, 48*(5), 799–801.

Eadie, P., Nguyen, C., Carlin, J., Bavin, E., Bretherton, L., & Reilly, S. (2014). Stability of language performance at 4 and 5 years: Measurement and participant variability. *International Journal of Language & Communication Disorders, 49*(2), 215–227. doi: 10.1111/1460-6984.12065

Eksborg, S. (1981). Evaluation of method-comparison data. *Clinical Chemistry, 27*(7), 1311–1312.

Eriksson, M., Westerlund, M., & Berglund, E. (2002). A screening version of the Swedish Communicative Development Inventories designed for use with 18-month-old children. *Journal of Speech, Language, and Hearing Research, 45*(5), 948–960. doi: 10.1044/1092-4388(2002/077)

Feldman, H., Dollaghan, C., Campbell, T., Kurs-Lasky, M., Janosky, J., & Paradise, J. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development, 71*(2), 310–322. doi: 10.1111/1467-8624.00146

Fiske, D. W. (1987). Construct invalidity comes from method effects. *Educational and Psychological Measurement, 47*(2), 285–307. doi: 10.1177/0013164487472001

Gabbe, B., Simpson, P., Sutherland, A., Palmer, C., Butt, W., Bevan, C., & Cameron, P. (2010). Agreement between parent and child report of health-related quality of life: Impact of time postinjury. *Journal of Trauma and Acute Care Surgery, 69*(6), 1578–1582. doi: 10.1097/TA.0b013e3181f8fd5f

Gardner, F. (2000). Methodological issues in the direct observation of parent–child interaction: Do observational findings reflect the natural behavior of participants? *Clinical Child and Family Psychology Review, 3*(3), 185–198. doi: 10.1023/A:1009503409699

Gilmore, L., & Cuskelly, M. (2011). Observational assessment and maternal reports of motivation in children and adolescents with Down Syndrome. *American Journal on Intellectual and Developmental Disabilities, 116*(2), 153–164. doi: 10.1352/1944-7558-116.2.153

Hawes, D., & Dadds, M. (2006). Assessing parenting practices through parent-report and direct observation during parent-training. *Journal of Child and Family Studies, 15*(5), 554–567. doi: 10.1007/s10826-006-9029-x

Hayden, E., Durbin, E., Klein, D., & Olino, T. (2010). Maternal personality influences the relationship between maternal reports and laboratory measures of child temperament. *Journal of Personality Assessment, 92*(6), 586–593. doi: 10.1080/00223891.2010.513308

Hopkins, W. (2004). Bias in Bland-Altman but not regression validity analyses. *Sportscience, 8*, 42–46.

Hudson, S., Levickis, P., Down, K., Nicholls, R., & Wake, M. (2015). Maternal responsiveness predicts child language at ages 3 and 4 in a community-based sample of slow-to-talk toddlers. *International Journal of Language & Communication Disorders, 50*(1), 136–142. doi: 10.1111/1460-6984.12129

Kirkwood, B. R., & Sterne, J. A. C. (2013). *Essential medical statistics* (2nd ed.). Malden, MA: Blackwell Sciences Ltd.

Kramer, S. F., Cumming, T., Churilov, L., & Bernhardt, J. (2013). Measuring activity levels at an acute stroke ward: Comparing observations to a device. *BioMed Research International, 2013*, 8. doi: 10.1155/2013/460482

Law, J., & Roy, P. (2008). Parental report of infant language skills: A review of the development and application of the Communicative Development Inventories. *Child and Adolescent Mental Health, 13*(4), 198–206. doi: 10.1111/j.1475-3588.2008.00503.x

Lee, J., Koh, D., & Ong, C. N. (1989). Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Computers in Biology and Medicine, 19*(1), 61–70. doi: 10.1016/0010-4825(89)90036-X

Lin, I. L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255–268. doi: 10.2307/2532051

Ludbrook, J. (1997). Comparing methods of measurements. *Clinical and Experimental Pharmacology and Physiology, 24*(2), 193–203.

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology, 29*(7), 527–536. doi: 10.1046/j.1440-1681.2002.03686.x

Ludbrook, J. (2010a). Confidence in Altman–Bland plots: A critical review of the method of differences. *Clinical and Experimental Pharmacology and Physiology, 37*(2), 143–149. doi: 10.1111/j.1440-1681.2009.05288.x

Ludbrook, J. (2010b). Linear regression analysis for comparing two measurers or methods of measurement: But which regression? *Clinical and Experimental Pharmacology and Physiology, 37*(7), 692–699. doi: 10.1111/j.1440-1681.2010.05376.x

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23*(1), 151–169. doi: 10.1146/annurev.publhealth.23.100901.140546

McBride, G. (2005). *A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient. NIWA Client Report: HAM2005-062.* Hamilton, New Zealand: National Institute of Water and Atmospheric Research Ltd. Retrieved from www.medcalc.org/download/pdf/McBride2005.pdf.

Miles, J., & Banyard, P. (2007). *Understanding and using statistics in psychology: A practical introduction.* London, UK: Sage Publications.

Mullineaux, D. R., Barnes, C. A., & Batterham, A. M. (1999). Assessment of bias in comparing measurements: A reliability example. *Measurement in Physical Education and Exercise Science, 3*(4), 195–205. doi: 10.1207/s15327841mpee0304_1

Nickerson, C. A. E. (1997). A note on 'a concordance correlation coefficient to evaluate reproducibility'. *Biometrics, 53*(4), 1503–1507. doi: 10.2307/2533516

Olino, T., Durbin, E., Klein, D., Hayden, E., & Dyson, M. (2013). Gender differences in young children's temperament traits: Comparisons across observational and parent-report methods. *Journal of Personality, 81*(2), 119–129. doi: 10.1111/jopy.12000

Pan, B., Rowe, M., Spier, E., & Tamis-LeMonda, C. (2004). Measuring productive vocabulary of toddlers in low-income families: Concurrent and predictive validity of three sources of data. *Journal of Child Language, 31*(3), 587–608. doi: 10.1017/S0305000904006270

Quercioli, C., Messina, G., Barbini, E., Carriero, G., Fanì, M., & Nante, N. (2009). Importance of sociodemographic and morbidity aspects in measuring health-related quality of life: Performances of three tools. *The European Journal of Health Economics, 10*(4), 389–397. doi: 10.1007/s10198-008-0139-9

Ring, E., & Fenson, L. (2000). The correspondence between parent report and child performance for receptive and expressive vocabulary beyond infancy. *First Language, 20*(59), 141–159. doi: 10.1177/014272370002005902

Sachse, S., & Von Suchodoletz, W. (2008). Early identification of language delay by direct language assessment or parent report? *Journal of Developmental & Behavioral Pediatrics, 29*(1), 34–41 doi: 10.1097/DBP.0b013e318146902a

Seifer, R., Sameroff, A., Barrett, L., & Krafchuk, E. (1994). Infant temperament measured by multiple observations and mother report. *Child Development, 65*(5), 1478–1490. doi: 10.1111/j.1467-8624.1994.tb00830.x

Zaslow, M., Weinfield, N., Gallagher, M., Hair, E., Ogawa, J., Egeland, B., ... De Temple, J. (2006). Longitudinal prediction of child outcomes from differing measures of parenting in a low-income sample. *Developmental Psychology, 42*(1), 27–37. doi: 10.1037/0012-1649.42.1.27