Age-related errors in the assessment of children

Scott Veldhuizen, MA

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Applied Health Sciences (Behavioural and Population Health)

Faculty of Applied Health Sciences
Brock University
St. Catharines, ON

© January, 2017
ABSTRACT

Children's functioning can only reasonably be measured relative to that of other individuals of the same age. In practice, age ranges are usually used to group children for this purpose. Examples include school grades, age groups in sports, and age bands used in developmental assessments. Age grouping is associated with systematic errors, often known as relative age effects (RAEs): Within each age group, older children outperform younger ones. This type of assessment error may lead to opportunities and interventions being offered inefficiently or unfairly.

This thesis comprises 5 research projects that aim to clarify underlying causes of RAEs, examine their importance in different contexts, and develop analytic methods relevant to their study. I use data drawn from two studies: A prospective cohort study including athletic performance measures (the Physical Health Activity Study Team project) and a validation study undertaken to compare measures of child development (Psychometric Assessment of the Nipissing District Developmental Screener). I develop linear models to characterize age-related variation and then use results to draw conclusions, to inform other analyses, and to generate synthetic datasets.

Together, studies demonstrate a set of methods for the exploration and correction of RAEs. They also yield several concrete findings: (1) A simple mathematical interpretation of RAEs can fully explain the errors seen in real datasets, meaning that other explanations are, in at least some contexts, unnecessary. (2) RAEs have different effects in ranking and selection contexts, with ranking errors largest among average
individuals but selection errors greatest when more extreme thresholds are used. (3) Age bands cause misclassification in measures of child development, and the error rate rises rapidly with the width of age bands used. (4) The use of different sets of age bands will prevent different assessments from agreeing closely. (5) Age grouping in developmental assessments will create an illusion of longitudinal instability. Finally, I demonstrate the use of alternative scoring approaches and discuss how these can reduce or eliminate errors related to RAEs.
Acknowledgements

I would like to acknowledge the support of my supervisor, Professor Terrance Wade, and of my committee members, Professors John Cairney, Jan Frijters, and Panagiota Klentrou.
# TABLE OF CONTENTS

CHAPTER 1: Introduction ........................................................................................................... 1

    Prevalence of age-grouping errors .................................................................................. 2
    Relative age effects: Theoretical and empirical research ................................................. 3
    Age banding errors: A simple view .................................................................................. 4
    Importance of age grouping errors .................................................................................. 6
    Proposed research ............................................................................................................ 7
    Projects and research aims .............................................................................................. 7
    Conclusion ....................................................................................................................... 15
    References ....................................................................................................................... 17

CHAPTER 2: Relative age effects in fitness testing in a general school sample: How relative are they? .................................................................................................................. 20

    Introduction ...................................................................................................................... 21
    Method ............................................................................................................................... 25
    Measures ........................................................................................................................... 26
    Statistical Analysis .......................................................................................................... 27
    Results ............................................................................................................................... 28
    Discussion ......................................................................................................................... 32
    Limitations ....................................................................................................................... 34
    Conclusion ....................................................................................................................... 35
    References ....................................................................................................................... 36

CHAPTER 3: When and for whom are relative age effects important? Evidence from a simple test of cardiorespiratory fitness. .................................................................................... 42

    Introduction ...................................................................................................................... 43
    Methods and analysis ...................................................................................................... 46
    Results ............................................................................................................................... 48
    Discussion ......................................................................................................................... 51
    Limitations ....................................................................................................................... 56
    Conclusion ....................................................................................................................... 57
    References ....................................................................................................................... 58

CHAPTER 4: Misclassification due to age grouping in measures of child development. 64

    Introduction ...................................................................................................................... 65
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Errors, misclassifications, and instability due to the use of age bands in measures of child development</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>Introduction</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>Methods</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Results</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Discussion</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>98</td>
</tr>
<tr>
<td>6</td>
<td>Concurrent validity of the Ages and Stages Questionnaires and Bayley developmental scales in a general population sample</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td>Introduction</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>Methods</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>Analysis</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>Results</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>Discussion</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td>Limitations</td>
<td>119</td>
</tr>
<tr>
<td></td>
<td>Conclusion</td>
<td>119</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>121</td>
</tr>
<tr>
<td>7</td>
<td>Summary, conclusions, and future research</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>Methods</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>Assumptions</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td>Summary of findings and commentary</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>Implications of age banding</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>Alternatives to age banding</td>
<td>144</td>
</tr>
<tr>
<td></td>
<td>Future research</td>
<td>146</td>
</tr>
<tr>
<td></td>
<td>Conclusion</td>
<td>149</td>
</tr>
</tbody>
</table>
### LIST OF TABLES AND FIGURES

<table>
<thead>
<tr>
<th>Table/Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2-1</td>
<td>PHAST cohort descriptives.</td>
<td>29</td>
</tr>
<tr>
<td>Table 2-2</td>
<td>Mixed-effects growth model for shuttle run performance among males in PHAST sample.</td>
<td>30</td>
</tr>
<tr>
<td>Table 2-3</td>
<td>Mixed-effects growth model for shuttle run performance among females in PHAST sample.</td>
<td>30</td>
</tr>
<tr>
<td>Figure 2-1</td>
<td>Illustration of relative age effects due to age differences alone.</td>
<td>31</td>
</tr>
<tr>
<td>Figure 2-2</td>
<td>Predicted Leger run stage by age for PHAST males and females with 95% confidence intervals.</td>
<td>32</td>
</tr>
<tr>
<td>Figure 3-1</td>
<td>Predicted mean and variance for Leger shuttle run performance by age.</td>
<td>49</td>
</tr>
<tr>
<td>Table 3-1</td>
<td>Predicted errors in percentile ranks for the Leger shuttle run test, by exact age and actual ability level.</td>
<td>50</td>
</tr>
<tr>
<td>Figure 3-2</td>
<td>Effect of performance cut-point on resulting age mixture (hypothetical distributions).</td>
<td>55</td>
</tr>
<tr>
<td>Figure 4-1</td>
<td>Illustration of misclassifications resulting from relative age effects.</td>
<td>66</td>
</tr>
<tr>
<td>Figure 4-2</td>
<td>Illustration of disagreement between instruments arising from use of different age groups.</td>
<td>67</td>
</tr>
<tr>
<td>Figure 4-3</td>
<td>BSID-III cognitive subscale raw scores by adjusted age from PANS study (n=593).</td>
<td>73</td>
</tr>
<tr>
<td>Figure 4-4</td>
<td>Example of generated data (n=593).</td>
<td>74</td>
</tr>
<tr>
<td>Table 4-1</td>
<td>Linear regression models for BSID-III cognitive subscale mean and variance.</td>
<td>74</td>
</tr>
<tr>
<td>Figure 5-1</td>
<td>Correlations among BSID-III subscales by age decile in PANS.</td>
<td>90</td>
</tr>
<tr>
<td>Figure 5-2</td>
<td>Proportion of cases misclassified as non-cases, by exact age, for narrow (BSID) age bands.</td>
<td>92</td>
</tr>
<tr>
<td>Figure 5-3</td>
<td>Variation in calculated percentile ranks due to age banding for a child who is average at all ages (‘cognitive’ BSID-III subscale).</td>
<td>93</td>
</tr>
<tr>
<td>Table 6-1</td>
<td>PANS sample descriptives (total n=587).</td>
<td>111</td>
</tr>
<tr>
<td>Figure 6-1</td>
<td>Derived -2SD threshold for BSID-III receptive communication subscale.</td>
<td>113</td>
</tr>
<tr>
<td>Table 6-2</td>
<td>Measures of agreement between ASQ and BSID-III in PANS.</td>
<td>114</td>
</tr>
</tbody>
</table>
CHAPTER 1: Introduction

Measurement of functioning is complex. Physical characteristics like height and T cell count represent physical realities that can be measured directly and, within technical limitations, accurately. Abilities, complex skills, and mental states present a more difficult problem. The thing to be measured must first be defined, the meaningfulness of this definition assessed, and a way of evaluating it devised. Measurement is indirect, and the result is often not a ratio or interval variable, but rather a location on a distribution.

Pediatric assessments present additional difficulties. Most of these arise from the fact that children are in a state of continuous, albeit nonlinear, development. A sample of adults can usually be meaningfully ranked, but children can be compared only to others of the same age. A height that would be average for a three-year-old would be unusually short for one who is four or five, and the same is true for measures of skills and abilities.

Efforts to measure ability in children must therefore take age into account. Although height and weight norms are usually presented as continuous growth curves, few other things are measured in this way. More common is some form of age banding, whereby age ranges are used to divide children into discrete groups and children within each group are then compared to one another or to a common standard. The most obvious example of age banded assessment is in schools, where children are grouped into grades for instruction and assessment. The same thing is done in sports and athletics, where age groups are used to ensure that individuals compete against others of approximately the same age.
Age banding is also ubiquitous, however, in instruments that seek to measure developmental functioning. Some, like the Nipissing District Developmental Screen and the Ages and Stages Questionnaires, provide a distinct set of milestones for each age group. Others, like the Bayley Scales of Infant and Child Development, provide a list of items ordered approximately by difficulty (i.e., a “power test”) that can be applied to all children, but make use of age bands to score these results.

**Prevalence of age-grouping errors**

The problem with age banding is that it systematically overestimates functioning among children who are chronologically old for their group and underestimates it for those who are young. In schools, the oldest children in the class will have a certain advantage, because of their greater experience levels and more advanced cognitive development. The same is true in athletic settings, with older children generally having better motor functioning and being larger and more physically mature. These are clear advantages, with skeletal maturation, for example, having been shown to be more advanced among athletes than other children of the same age (Moore et al., 2010). Exactly the same issue applies in developmental testing, where children towards the older limit of an age band will tend to outperform younger ones.

In sport, athletic, and educational research, these and related problems are known as *relative age effects* (RAE). Considerable effort has been expended to measure the advantages and disadvantages created by age gradients in this context. Decades of
research have shown, for example, that elite sport and selective universities contain a disproportionate number of people born early in the year (e.g., Cobley, Baker, Wattie, & McKenna, 2009; Cobley, McKenna, Baker, & Wattie, 2009; Deprez, Vaeyens, Coutts, Lenoir, & Philippaerts, 2012; Musch & Grondin, 2001; Roberts, Boddy, Fairclough, & Stratton, 2012; Sherar, Baxter-Jones, Faulkner, & Russell, 2007). Conversely, children who are younger than most of their classmates are considerably more likely than others to be identified as having behavioral problems (Morrow et al., 2012). RAEs in athletics have also been shown to exist among older adults, at age ranges where performance declines rather than improves (e.g. Medic, Starkes, Weir, Young, & Grove, 2009).

Relative age effects: Theoretical and empirical research

The first studies in this area were empirical. RAEs came to be widely-studied in sport when it was noticed that professional hockey players were more likely to have been born in January than December, and had therefore been older than most team-mates throughout their childhood and adolescence (Barnsley et al., 1985). This was followed by a wave of studies documenting the presence, or occasionally the absence, of RAEs in various settings. A smaller set of papers speculated about causes. Some early work considered the possibility that teaching or coaching effects were partly responsible. This might take the form of “Pygmalion” effects, in which performance is affected by higher expectations, or of “Golem” effects, in which individual performance falls because expectations are low (Babad, Inbar, & Rosenthal, 1982; Hancock, Adler, & Côté, 2013). Other researchers mooted the possibility of season-of-birth effects (reviewed in Musch and Grondin, 2001) of the kind that have occasionally been reported for neurological and
psychiatric conditions, notably schizophrenia (Torrey, Miller, Rawlings, & Yolken, 1997). In this view, children born in winter may have an increased exposure to infectious disease and fever, or lower availability of essential chemicals such as vitamin D. In more recent years, attention has focused on the simpler and more clearly plausible role of biological maturation. In the maturation-selection hypothesis, which has been applied largely to elite sport, relatively old children are more likely to be selected to participate, and therefore have advantages that lead to their overrepresentation in professional sport later in life (Cobley et al., 2009a; Cobley et al., 2009b).

Much less work has been done on age banding effects in other contexts. The term “relative age effect” is largely restricted to research on sports and athletics. Under different names, however, the problem is familiar to developers, and some users, of developmental assessments (Boyd, 1989), and, as noted earlier, research has shown that relative age is associated with the identification of behavioral problems (Morrow et al., 2012). Research in the fields of child development and psychiatry has tended to allude to the problem or, sporadically, to identify it empirically in a specific situation. It has received little or no attention as a source of systematic error.

**Age banding errors: A simple view**

As noted, existing work on age banding effects has focused on a variety of possible causes. This thesis begins from the premise that banding errors in all settings are the mathematically inevitable result of age differences within groups. It is straightforward to show that age-related errors will be present whenever (a) systematic age-related variation
is present, and (b) people within an age range are compared to each other or to a common
standard. These errors can be expected to be ubiquitous. The existence of age groups, in
fact, signals that such errors are present: If ability did not vary systematically with age,
then age groups would not be necessary; and if age groups are present and ability varies
with age, then there will be systematic variation within each group. Curiously, this
simple fact has not been clearly stated in the literature.

From this perspective, many posited explanations are potentially superfluous, and all are
probably incomplete. To approach the problem by positing and testing individual causes
is, in fact, to some extent to miss the point. Age-related errors like RAES must reflect the
sum of all factors that cause age-related variation in ability. Even biological maturation
is clearly an incomplete explanation. In sport, greater age will confer advantages in terms
of size and motor ability; but it also usually means greater experience, different social
expectations, and so on. In developmental assessments, brain development is crucial, but
variation will also reflect differences in education received or life experience more
generally. The reasons for RAES in most contexts are simply the reasons that age is
related to performance.

It is important to acknowledge that some age-related variation is more complex than this,
because advantages from initial classification errors can accrue over time. University
students born early in the year, for example, probably enjoy only a very modest
advantage from their greater age; instead, they continue to benefit from the good grades,
scholarships, confidence, and perhaps additional instruction they received earlier in life.
The same explanation applies to age differences in elite adult sport. In childhood, however, we are usually concerned with the immediate advantage conferred by relatively greater age. This is the issue that applies to the important questions of academic fairness and developmental assessment.

**Importance of age grouping errors**

Age banding produces errors in both ranking and classification, and all individuals will be affected to some degree. Grades and other assessments will be biased to a greater or lesser degree depending on the relative age of each child. This is a particular concern when selections are made, as some children will be wrongly identified as having either exceptionally poor or exceptionally good performance. Extra assistance and special opportunities will therefore not be distributed fairly or optimally: Some individuals who would benefit from this attention will not receive it.

RAEs will also produce misclassifications in developmental assessments. Younger children will be wrongly identified as delayed (i.e., false positives), and older ones with genuine delays will sometimes not be diagnosed (i.e., false negatives). As in academic and athletic settings, children who would benefit from support will not receive it, while others will be wrongly identified as “delayed”. The potential harms are considerable. Delayed children who do not receive appropriate support in early childhood may experience lifelong consequences. Conversely, wrongly identifying children as delayed means that clinical resources will be misdirected, that parents will experience unnecessary anxiety, and that the child may ultimately suffer from lowered expectations.
No existing work has systematically explored this kind of misclassification in developmental assessments, and theoretical work in sport and athletic settings has been limited.

**Proposed research**

This research program aims to clarify underlying causes of RAEs, to examine their importance in different contexts, and to develop analytic methods relevant to their study. This work received ethics approval from the Brock University Research Ethics Board (#15-266-WADE).

Project 1 presents a simple theory of RAEs, and demonstrates a method by which the usefulness of more elaborate explanations can be tested. Project 2 extends this work to make an important distinction between two general contexts in which RAEs can occur, and to demonstrate their relative importance in each. Project 3 measures RAEs in developmental assessments, and shows that they lead to misdiagnosis of developmental delay and to systematic disagreement between measures. Project 4 extends the work in Project 3 to more complex measures, and demonstrates the existence of other kinds of systematic error related to age in pediatric assessments. Finally, project 5 proposes an alternative to the use of age groups suitable to general-population samples and applies it to a real-world assessment study.

**Projects and research aims**

1) Relative age effects in fitness testing in a general school sample: how relative are they?
As noted, many explanations for RAEs have been proposed, but it is not clear that any are necessary. In this project, I make a distinction between artifacts caused by the imposition of age groups, which are probably inevitable, and truly relative effects, i.e., differences in performance associated with age relative to peers.

I examine this question using data on a simple athletic task: A 20m shuttle run test repeated twice a year for 4 years by a sample of approximately 1000 male students aged 9 or 10 years at baseline. To characterize age-related changes in performance, I fit a polynomial growth curve model. To search for an effect for age relative to peers, I then analyze this model’s residuals. If there are within-year effects, caused by grade effects, these would appear as a gradient over day of birth, with participants born early in the year having better-than-expected performance.

Project aims:

a) Argue that RAEs are a mathematical inevitability in all situations involving the application of common standards to people of different ages when the ability being measured varies systematically with age. RAEs are likely to be almost universal in educational, athletic, sport, research, and clinical settings involving children and adolescents.

b) Present a method for establishing whether RAEs in a given setting arise from simple age differences or whether any other factors play a role.

c) Demonstrate this method in a real dataset.
2) When and for whom are relative age effects important? Evidence from a simple test of cardiorespiratory fitness.

RAEs must be ubiquitous, but they have been studied mostly in ‘selection’ contexts – situations in which people are identified as exceptional. Analyses of elite sport participation, for example, typically measure age gradients among people who have previously been selected for exceptional performance: A disproportionate number of athletes with birthdays in January implies a selection from within groups formed by year of birth. However, RAEs also confer advantages and disadvantages in more general ranking situations – and these are arguably more important, as they affect all individuals, not just those who are exceptional.

In academic contexts, older children within each grade will tend to have an advantage due to their greater experience and more advanced development. Although the underlying process producing advantages is the same in both selection and ranking contexts, the observed pattern of results can be expected to differ for statistical reasons. In this project, I make a distinction between these two circumstances, using the same dataset, and the same growth curve model results, as project 1.

In this analysis, I fit another model examining the change in variance with age. With equations describing how age affects both the mean and variance of performance, I then develop arguments showing how the importance of RAEs differs in grading and selection contexts. In particular, I show that, in ranking terms, RAE-related errors are greatest in
the middle of the distribution of performance, but that the size of RAEs observed after a selection is made increase as the selection threshold becomes more extreme. I also note that RAEs can, for a known distribution of performances, be derived from a simple formula.

Aims:

a) Distinguish between ranking and selection contexts, and explain the functioning of RAEs in each.

b) Examine the statistics of RAEs, and demonstrate that they produce very different effects in each of these contexts.

c) Consider the implications of this finding in each context (e.g., academic grading v. selection for elite sport).

3) Misclassification due to age grouping in measures of child development.
Although most work on RAEs has been conducted in sport and athletics, their implications are arguably more important in other contexts. In particular, age grouping is an important source of errors in pediatric assessments. Although this issue has occasionally been acknowledged in the literature in specific contexts, the underlying statistical issues have not been described, and no rigorous attempts have been made to systematically measure misclassifications or other errors.

A further and less obvious issue related to the use of age bands in developmental assessment is the fact that the use of different age groups by different measures can be
expected produce systematic disagreements between them. This may complicate clinical assessments in which multiple measures are used, and may also help to explain the relatively poor agreement observed in validation studies.

In this project, I use data from the Psychometric Assessment of the Nipissing District Developmental Screener (PANS) study, which administered a widely-used measure of child development (the third edition of the Bayley Scales of Infant and Child Development; BSID-III) to 594 children aged 1 month to 3 years. For the purposes of demonstration, I use the cognitive subscale of the BSID-III.

To measure misclassifications, I use simulation methods. To inform this analysis, I fit regression models describing age-related change in subscale scores and in the variance of subscale scores. I then use these equations to generate a simulated dataset of 1M “children”. Imposing the BSID-III’s age groups on this dataset then makes it possible to measure the numbers of false positives and false negatives. Imposing age bands from multiple instruments similarly makes it possible to estimate the maximum level of agreement to be expected from instruments using different age groups.

Aims:

a) Measure the misclassifications that occur due to age grouping using one subscale of a commonly-used measures of child development.

b) Demonstrate that age grouping also prevents different instruments from agreeing closely.
c) Demonstrate that age grouping is responsible for 1) a relatively high number of false positive and false negative results in screening and evaluation, and 2) the low measured agreement of different instruments.

d) Discuss alternative approaches to scoring that have the potential to avoid banding-related errors.

4) Errors, misclassifications, and instability due to the use of age bands in measures of child development.

Study 3 addresses a number of issues related to age banding, and develops a method for the measurement of errors, but relies for its demonstration on a single subscale of a single measure. Child development, however, consists of advances and skill acquisition across multiple related but distinct domains, including cognitive functioning, language use, and motor skills. As a result, most clinician-administered instruments attempt to measure functioning in multiple areas. Although the results in Study 3 illustrates the problem, a more complete evaluation of errors related to age banding requires an examination of an instrument with multiple subscales.

Study 3 also describes only errors in a single assessment. Another issue concerns repeated assessments. Age-related errors limit longitudinal stability, because the error for an individual child will differ at different ages. A child who is at the older end of an age group at one assessment and at the younger end at another, for example, will appear to experience a relative developmental slowing – but this is a statistical artifact caused by age banding. The result is likely to be a false appearance of instability, of discontinuous
development, or even of developmental regression. Study 3 also largely measures misclassification only; it does not thoroughly describe the size of errors, or their variation with exact age. These can be expected to vary sharply with age: The risk of misclassification, for example, is likely to be modest for children close to the centre of their age bands, but may become extreme for those close to a division between age bands.

In this study, I expand the analysis of Study 3 to examine these issues in a multi-domain instrument. I again use a combination of linear modeling and simulation to measure error rates. This analysis is complicated not only by the need to model expected values and variances as continuous functions of age for multiple domains, but also by the need to take into account inter-domain correlations; and, in particular, changes in this correlation structure with child age. To obtain mean and variance functions, I follow the same approach as in Study 3, but fit these models for all 5 domains of the BSID-III. I also measure inter-domain correlations within age subsets of the PANS dataset, and use the results to derive estimates of each correlation as a linear function of age. I then use these results to develop a large synthetic dataset. An exploration of this dataset then makes it possible to explore age-related errors of different kinds.

Aims:

a) Extend the analysis in paper 3 to the case of complete developmental assessments, which commonly assess multiple domains of functioning,

b) More precisely measure the probability of misclassification at different child ages.
c) Demonstrate that RAEs and other age effects lead to unstable longitudinal courses and limit predictive validity.

d) Demonstrate that age-related changes in the associations between different developmental domain produce systematic differences in prevalence with age.

5) Concurrent validity of the Ages and Stages Questionnaires and Bayley Developmental Scales in a General Population Sample.

Earlier projects demonstrate that scoring developmental assessments using age groups is problematic, and develop methods for measuring the errors involved, but do not provide an alternative.

In this study, I undertake a validation study that supplements conventional results with methods that use internal rankings instead of supplied population norms – an approach that makes it possible to discard age bands altogether. The approach relies, as do previous studies, on the accurate modeling of the mean and variance of the score on developmental subscales as functions of age. Instead of proceeding to simulation, however, I subtract observed from predicted values for each study participant to produce standard scores of functioning that are fully adjusted for child age and that reflect each individual’s location on a distribution of functioning. This approach is only appropriate for large, general-population samples, and reduces the comparability of results to other samples; but also has the possible advantage of avoiding the use of norms that may be inappropriate or problematic. Related methods could, moreover, be used by instrument developers to produce scoring approaches that avoid age banding altogether.
Aims:

a) Apply arguments and findings to a real-world validation of a developmental instrument.

b) Present an analysis that ignores defined age groups in the reference measure and produces a location on the sample distribution for the exact age of each participant.

Conclusion

This thesis is a body of work that aims to progress from an elucidation of problems to a development of solutions. It begins by noting the universality and inevitability of errors produced by RAEs. I argue that these errors arise from simple statistical artifacts produced by the quantization of continuous data. Although RAEs have been reported and discussed within specific subject areas, they occur in many disciplines; and although their cause is easily understood, theoretical work has sometimes sought unnecessarily complex explanations while neglecting simple ones.

The proposed studies concern RAEs in athletics and in developmental assessment, but the approaches used are applicable in many other settings. Although the theory of RAEs is developed, the bulk of the work incorporated into this thesis is concrete. I use both actual and simulated data to develop methods of measuring errors in various situations, and highlight several issues that have not previously been discussed, including problems of predictive validity and the variation in expected errors across levels of functioning and
across contexts. Finally, I make an effort to develop a relatively tractable alternative that should be appropriate to at least some settings in which RAEs occur. The ultimate aims of this body of research are to develop a more complete conceptual and statistical understanding of age banding errors, to underscore their importance and universality, and to demonstrate approaches that can mitigate them.
References


CHAPTER 2: Relative age effects in fitness testing in a general school sample: How relative are they?

Scott Veldhuizen, MA\textsuperscript{1,2,3}, John Cairney, PhD\textsuperscript{2,3}, John A. Hay, PhD\textsuperscript{3}, Brent E. Faught, PhD\textsuperscript{3}

\textsuperscript{1}Centre for Addictions and Mental Health, Toronto, Ontario, \textsuperscript{2}Department of Family Medicine, McMaster University, Hamilton Ontario, \textsuperscript{3}Department of Health Sciences, Brock University, St. Catharines, Ontario

Introduction

The grouping of children and adolescents into groups based on age, as is done when they are divided into grades in school or into age bands in organized sport, produces groups with a mix of older and younger individuals. When children are enrolled in grades according to year of birth, for example, age differences of up to a year will be present. When these individuals compete against one another, either explicitly (as in athletic competitions) or implicitly (when graded relative to peers or norms in academics), older individuals have generally been shown to enjoy an advantage (Cobley et al., 2009a; Cobley et al., 2009b; Musch & Grondin, 2001; Roberts et al., 2012; Deprez et al., 2012; Sherar et al., 2007).

Differences of this kind are known as relative age effects (RAEs) (Musch & Grondin, 2001; Wattie et al., 2008). RAEs have been shown to play an important role in sport, with adolescent athletes who are old for their group consistently overrepresented at elite levels (Cobley et al., 2009a). The same effects have been shown to exist among older adults, at age ranges where performance declines rather than improves (e.g., Medic et al., 2009). RAEs also influence academic outcomes, including the identification of children as gifted or delayed (e.g., Cobley et al., 2009b, Dhuey & Lipscomb, 2010) and even the diagnosis of conditions such as attention deficit hyperactivity disorder (Morrow et al., 2012).

RAEs can be expected whenever a) age-related variation is present, and b) individuals are grouped by age range. The presence of such age groups should, in fact, probably be
taken as a sign that RAEs are present: If age conferred no advantage, age groups would not be necessary; if age groups are used, then older and younger individuals are being mixed together, and one group or the other will enjoy a systematic advantage.

In some settings, such as school classes, RAEs are likely to take the form of systematic differences in performance. When individuals choose or are chosen to participate (or when they are identified as “delayed” or “gifted”), on the other hand, RAEs are likely to take the form of selection effects: The distribution of ages within a given group will show a gradient reflecting the advantages or disadvantages associated with being older or younger. In organized sport, especially, individuals who are old for their age group are more likely to participate (Musch & Grondin, 2001; Baker et al., 2010). In academics, children who are young for their grade may be absent more often (Cobley et al., 2009b), and may be less likely to ultimately pursue post-secondary education (Bedard & Dhuey, 2006). It has also been noted that age relative to peers could affect the probability of being required to repeat a grade or the decision to defer school entry (Dhuey & Lipscomb, 2008).

Most analyses of RAEs have demonstrated their existence by showing an age gradient within groups in performance or participation. If individuals are grouped by year of birth, for example, individuals born between January and March may be compared to those born between October and December. If performances differ between these groups, or if one group is overrepresented, RAEs are judged to be present, and attention shifts to considerations of why they exist. Apart from maturational differences, several
explanations have been put forward. Being consistently older and stronger than peers might improve performance further because confidence or past success encourage increased training or practice (Cobley et al., 2009), and/or through “Pygmalion” effects, in which expectations affect performance (Babad et al., 1982; Hancock et al., 2013). The possibility of season-of-birth effects has also been raised (reviewed in Musch & Grondin, 2001): Children born in winter may be young or old relative to peers (or a mix of both extremes), and this may be associated with performance through season of birth effects resulting from gestational factors or early childhood experiences (such effects, albeit weak and debated ones, have been reported for certain neurological and psychiatric conditions, notably schizophrenia (Torrey et al., 1997)).

RAEs can also, however, be understood as arising naturally and straightforwardly from an underlying relationship between age and performance. Figure 1 illustrates this effect. The dashed line represents the average level of ability by child age, and the horizontal lines represent the measured average within each grade. The dots represent two students, one born near the beginning of the year, one near the end. The older child appears markedly above average, the younger one below; but each is, in fact, exactly average for his or her age. In this example, RAEs are simply age effects; they come about because we have imposed categories on something that varies continuously with age. Older children within a grade outperform younger ones, on average, for the same reasons later grades, on average, will outperform earlier ones. This can be presumed to be due to some combination of maturation and accrued experience, but the essential point is that there is
nothing specifically relative about the phenomena at work: The question is simply why people get better with age on the task being considered.

This distinction has not, to our knowledge, been considered in existing work. A reliance on cross-sectional data and on quantile-based analysis may therefore have caused explanations to multiply unnecessarily. In this article, we consider a simple fitness task, and use longitudinal data and growth models to a) identify groups for whom RAEs are likely to be most important, and b) test whether RAEs result simply from age-related improvement, or whether other factors (such as age relative to peers) are operating. We argue that, if other factors are operating, then there will be systematic differences after age and sex are fully accounted for. A lack of such variation, meanwhile, would suggest that the factors that explain differences within age groups are the same as those that explain differences between them, and that – for this task, in this population – it is not necessary to invoke other explanations.

To do this analysis, we use a large, longitudinal dataset of Leger shuttle run performances. The shuttle run is a measure of cardiorespiratory fitness (Leger & Lambert, 1982) for which RAEs have previously been identified (Roberts et al., 2012). The shuttle run is of considerable interest, because it is widely used throughout North America as an assessment of fitness (Ernst et al., 2006; Lobello et al., 2009). In school-based settings, it is common practice to administer and report the test at the grade level, which means that RAEs may influence results (Roberts et al., 2012). The shuttle run also provides a simple measure of maximal running speed, which will be directly related to
performance in sport or athletic tasks involving sprinting (e.g., track events, soccer, baseball).

Method
Data for this study come from the Physical Health Activity Study Team (PHAST) project, a large, longitudinal study of health, motor functioning and associated characteristics conducted between September 2004 and June 2009 (Cairney et al., 2012). The target population included all children enrolled in Grade 4 in the public-school system of a large region of Southern Ontario, Canada. Ethics approval was obtained from both the local District School Board and Brock University. In year 1 (September 2004 to June 2005) of the study, we received permission from 75 of 92 schools (83%) and informed consent from the parents of 2278 of 2378 children (95.8%).

Testing protocols were established, and baseline pilot testing was completed in September and October 2004. Data collection began in the spring of 2005 (mean age 9.8 years, SD = 0.35) and ended in the fall of 2009 when the children were completing Grade 9 (mean age 14.6 years, SD = 0.36). For the first 3 years of the study (spring 2005 to spring 2007), data were collected twice a year in the fall and in the spring. Thereafter, data were collected once a year in the fall (fall 2007 to fall 2009). Children who enrolled in participating schools after baseline were included in the sample. Children enrolled in special education classes were excluded from the analysis.
The combination of enrolled students and testing periods yields 21,138 possible assessments. This does not correspond to the product of the number of waves and the number of subjects because PHAST was an open cohort; students entered and left eligible schools during the study. 17,923 (85%) measures were actually taken. Of these records, 15,240 (72%) had consent for data use, completed the Leger test, and were not in special-needs classes. The number of individuals with complete data fell in later waves, due largely to movements of students between classes and schools.

**Measures**

The Leger 20m Shuttle Run test (Leger & Lambert, 1982; Leger et al., 1998) is a measure used to estimate peak VO\(2\). It has been shown to provide reliable and valid estimates in children and adolescents (Leger & Gadoury, 1989), and is widely used, especially in school-based research and assessment. It and similar instruments (e.g., Cooper Institute for Aerobics Research, 2004) have been used internationally to identify and measure secular declines in levels of physical fitness (Kuntzleman & Reiff, 1992; Stratton et al., 2007; Tremblay et al., 2010).

When performing the shuttle run, participants run back and forth between two lines set 20m apart in synchrony to a beep emitted from an audio compact disc according to previously published procedures (Leger & Gadoury, 1989). In the PHAST study, subjects were tested in mixed-sex groups of 5 to 10. The test was complete when the child was unable to maintain the prescribed pace for two consecutive sound signals. The stage last attained before completion is used to estimate peak VO\(2\) (Leger & Gadoury, 1989); in this
analysis, however, we use it as our dependent measure directly.

We calculated age from date of birth and date of testing, and used this exact age in all analyses. Date of testing was not recorded in 2403 (15.8%) cases. In 221 (1.5%) cases, the test date could be inferred from those of classmates. In the remaining 2182 (14.3%), we used the average testing date for that wave. Testing was performed within an approximately one-month period at each wave, and the expected errors are therefore small. We calculated age relative to peers by subtracting the mean age for the child’s class at each assessment from the child’s own age. We did not use the grand mean for the child’s grade because this seems less relevant to issues of comparative performance: Children will be compared, and will compare themselves, to their immediate peers.

**Statistical Analysis**

We used mixed effects growth models (Singer & Willett, 2003), with a random effect for child age and separate random intercepts at the school level (due to cross-classification resulting from movements of children between schools). This approach takes into account the nesting of observations within children, and of children within schools. We omitted class-level effects from the models, on grounds that fitting cross-classified models of this complexity can become problematic, and that variance-components models for individual waves suggested that class-level variance was modest. Due to known sex differences in the relationship between age and performance (confirmed in this sample by testing an interaction), we divided the sample and analyzed males and females separately.
We expected the association between age and performance to be non-linear. As capturing this association accurately is crucial to the measurement of relative age effects and to the testing of our research questions, we used the method of fractional polynomials. We fit models with 1 or 2 age terms and selected the best-performing transformations for each sex (Models 1a and 1b). For males, a single quadratic term provided the optimal solution. For girls, the association was non-monotonic, and we included both quadratic and cubic terms.

We tested for effects of age relative to peers by subtracting the child age from the mean age of their class at the time of assessment and including this term in the growth models. We also examined the associations between this relative-age term and the residuals from models 1a and 1b, again using the method of fractional polynomials to identify potential non-linear effects. Finally, we examined the Model 1a and 1b residuals by month of birth, and by fitting models to search for any linear or non-linear association with age relative to peers. We used Stata 13 for all analyses (StataCorp, 2013).

**Results**

Sample descriptives are shown in Table 1. For males, growth models (Table 2) showed a strong, consistent association between age and performance. For females (Table 3), performance increases were marked at young ages, but ceased at approximately age 13, with performance thereafter stable or slightly declining. Predicted performance by age is illustrated in Figure 2; the vertical difference between adjacent dots corresponds to the expected difference for a one-year age gap.
Table 2-1. PHAST cohort descriptives.

<table>
<thead>
<tr>
<th>Time point</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Age (mean; SD)</td>
</tr>
<tr>
<td>1</td>
<td>1127</td>
<td>9.9 (0.4)</td>
</tr>
<tr>
<td>2</td>
<td>1073</td>
<td>10.3 (0.4)</td>
</tr>
<tr>
<td>3</td>
<td>1067</td>
<td>10.8 (0.5)</td>
</tr>
<tr>
<td>4</td>
<td>1041</td>
<td>11.4 (0.4)</td>
</tr>
<tr>
<td>5</td>
<td>1020</td>
<td>11.9 (0.4)</td>
</tr>
<tr>
<td>6</td>
<td>893</td>
<td>12.4 (0.3)</td>
</tr>
<tr>
<td>7</td>
<td>833</td>
<td>13.4 (0.3)</td>
</tr>
<tr>
<td>8</td>
<td>678</td>
<td>14.6 (0.4)</td>
</tr>
</tbody>
</table>

These results provide measures of RAEs resulting directly from age differences. Among males, the oldest children in a grade would generally be expected to outperform the youngest by 0.26 Leger stages at ages 9/10, and by 0.6 stages at ages 14/15. Larger differences correspond to higher overall levels of performance, however, which means that differences rise more slowly in percentage terms: The performance gap is expected to be 7% at ages 9/10 and 10% at ages 14/15. Among females, performance increases with age slow and ultimately cease, and the largest RAEs can therefore be expected at younger ages. At ages 9/10, the expected gap is 0.46 stages (15% of average performance), while the variation from 12 to 15 is negligible. The gap between males and females therefore first narrows, reaching a minimum at age 11, and then widens rapidly after age 12.
Table 2-2. Mixed-effects growth model for shuttle run performance among males in PHAST sample.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>z</th>
<th>P</th>
<th>L95%</th>
<th>U95%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age(^2)</td>
<td>0.017</td>
<td>0.00068</td>
<td>24.86</td>
<td>&lt;0.001</td>
<td>0.016</td>
<td>0.0182</td>
</tr>
<tr>
<td>Intercept</td>
<td>2.06</td>
<td>0.11</td>
<td>18.82</td>
<td>&lt;0.001</td>
<td>1.85</td>
<td>2.28</td>
</tr>
<tr>
<td><strong>Random effects (variances)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td>0.2</td>
<td>0.04</td>
<td>0.13</td>
<td>0.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age(^2)</td>
<td>0.0002</td>
<td>0.000017</td>
<td>0.00017</td>
<td>0.00024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>4.59</td>
<td>0.37</td>
<td>3.9</td>
<td>5.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cov(age, intercept)</td>
<td>-0.019</td>
<td>0.0023</td>
<td>-0.024</td>
<td>-0.015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>1.09</td>
<td>0.02</td>
<td>1.05</td>
<td>1.13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2-3. Mixed-effects growth model for shuttle run performance among females in PHAST sample.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>z</th>
<th>P</th>
<th>L95%</th>
<th>U95%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age(^2)</td>
<td>0.080</td>
<td>0.0064</td>
<td>12.43</td>
<td>&lt;0.001</td>
<td>0.067</td>
<td>0.093</td>
</tr>
<tr>
<td>Age(^3)</td>
<td>-0.0039</td>
<td>0.00036</td>
<td>-10.74</td>
<td>&lt;0.001</td>
<td>-0.005</td>
<td>-0.003</td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.12</td>
<td>0.29</td>
<td>-3.83</td>
<td>&lt;0.001</td>
<td>-1.686</td>
<td>-0.545</td>
</tr>
<tr>
<td><strong>Random effects (variances)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td>0.11</td>
<td>0.02</td>
<td>0.07</td>
<td>0.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.06</td>
<td>0.01</td>
<td>0.05</td>
<td>0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>7.38</td>
<td>0.74</td>
<td>6.07</td>
<td>8.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cov(age, intercept)</td>
<td>-0.61</td>
<td>0.06</td>
<td>-0.74</td>
<td>-0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>0.72</td>
<td>0.01</td>
<td>0.69</td>
<td>0.75</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other results concern the question of whether relative differences are fully accounted for by age alone. The age-for-class term (i.e., the difference between child age and the mean age of the class) was not significant when added to either model, suggesting that performance was not affected by age relative to immediate peers. Similarly, we found no significant association between day of birth (i.e., the day of the year, with January 1 as 1) and the outcome. We also tested for linear or non-linear associations between these terms and the residuals from models 1a and 1b. Means by month of birth showed no particular
pattern, and no monotonic associations were present between residuals and either age relative to peers or day of birth.

Figure 2-1. Illustration of relative age effects due to age differences alone.
Discussion

This study had two aims: To measure the size of RAEs for a fitness test in a general-population sample, and to determine whether RAEs were driven simply by age or whether other factors were important.

Results indicate that, in this context, RAEs result simply from age differences. We characterized the relationship between age and performance, and found no important further variation with age relative to peers or season of birth. This implies that, for this population and activity, relative age effects can be fully predicted from the general association between age and performance: Maturation, or other age-related differences
(such as accrued experience) are responsible for variation within age groups, and RAEs can be understood simply as the advantage of older individuals over younger ones. Results do not support differences due to season of birth, or to motivational, coaching, or other factors. The absence of any variation by month or season of birth agrees with most existing work (Musch & Grondin, 2001), but our results also imply that psychological, social, or other factors associated with being old or young relative to peers are not important for this type of fitness task in this population. Instead, RAEs arise simply from the underlying relationship between age and performance: Older children outperform younger ones, presumably due to maturation or accrued experience. The factors that explain age-related differences within grades are the same as those that explain differences between grades, and RAEs in this context simply result from the grouping of individuals by age.

**Shuttle run**

With this type of variation generally excluded, we can consider the size of RAEs as a function of the underlying association between age and performance. In our data, performance among males increased steadily with age, which means that RAEs are present for all ages represented in our sample. Our models also indicate that RAEs, in terms of shuttle run stage completed, will tend to be somewhat more important among older adolescents. This is the result of the generally accelerating improvements with increased age.

Age-related increases in performance were much more modest for females. This is
generally unsurprising, given well-established differences in the biological processes of maturation in each sex, as well as the general decrease in participation widely reported for adolescent girls (Pate et al., 1994). This means that, in general population settings, RAEs among girls are likely to be largest at younger ages: In our models, the largest advantages and disadvantages are for girls under age 12, and there will be very small effects for older individuals. These results broadly agree with those reported in a recent meta-analysis (Cobley et al., 2009a) of relative age effects in sport, which found that these were most pronounced among males aged 15 to 18.

**Limitations**

It is important to note that our results apply to a certain type of athletic performance in a general-population sample. They are not necessarily generalizable to organized athletics, to academics, or to other types of performance. Selection effects, in particular, are likely to be muted in our sample. A common finding in sport is that individuals born towards the end of the year (when birth year is used to define groupings) are overrepresented at elite levels. Elite sport, however, involves selection of a tiny proportion of players, while our data covers a general school sample at ages when rates of school leaving are very low (Gilmore & McKullen, 2012).

Another consideration concerns our outcome measure. Leger run stages are clearly a valid ordinal measure, in that completing more stages corresponds to greater ability, but may not be linearly related to the underlying “athletic performance” trait we might want to measure. If this is of greater interest than the final shuttle run stage completed, then
differences in absolute and percentage terms should be regarded as approximate. The
general pattern of results, however, should be unaffected.

**Conclusion**

We found no evidence of effects for age relative to peers or of season-of-birth effects.
This implies that, for cardiorespiratory fitness in the general school population of children
and adolescents, relative age effects are simply age effects, and that it is unnecessary to
consider explanations concerning confidence, motivation, selection, or maturation. It
also appears that the importance of RAEs for any particular age/sex mix can be estimated
directly; this may have implications for the types of groupings used (e.g., sex separation,
width of age bands), or for the evaluation of individuals within these groups. In
particular, methods of grading or ranking that consider an individual’s exact age may –
while perhaps difficult to implement in some settings – have the potential to mitigate
RAEs.
References


StataCorp (2013). Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.


CHAPTER 3: When and for whom are relative age effects important? Evidence from a simple test of cardiorespiratory fitness.

Scott Veldhuizen, MA¹,²,³, Terrance J. Wade, PhD³, John Cairney, PhD²,³, John A. Hay, PhD³, Brent E. Faught, PhD³

¹Centre for Addictions and Mental Health, Toronto, Ontario, ²Department of Family Medicine, McMaster University, Hamilton Ontario, ³Department of Health Sciences, Brock University, St. Catharines, Ontario

Introduction

Measures of child development or performance are typically interpreted through a comparison to peers. This may take the form of direct comparisons among children of the same gender and similar age, or may be done by comparison of performances to age-appropriate standards developed from a normative sample. In formal screens for developmental delay and in assessments for high or low functioning in education or athletics, it is common practice to group children by age and apply decision rules derived ultimately from a normative sample. In athletics, children are similarly grouped by age and then assessed against each other or against a common standard.

This pooling of individuals across an age range creates one or two important problems. In school settings and organized sport, it is common to group children by chronological annual age-groups. A year is a long time in childhood and adolescence, and, within a grade, there will be systematic differences in development and accrued experience between individuals of different ages. In general, the oldest individuals in each group enjoy considerable advantages due to their greater experience and maturity, and this is reflected in better performance in athletics and academics (see Musch et al., 2001; Wattie et al., 2008; Cobley et al., 2009A). Phenomena of this type are known as relative age effects (RAEs), and they should be expected whenever age and ability are related and people of different ages are grouped together.

Many mechanisms underlying RAEs have been proposed (reviewed in Musch et al., 2001). Roles have been proposed for teacher or coach interactions, expectations, access
to training, methodological biases (Delorme et al., 2010), season of birth, and the compounding of advantage from year to year (mechanisms are reviewed in Musch et al., 2001; Wattie et al., 2008; Cobley et al., 2009). RAEs have also been shown to persist into adulthood, after any maturational advantage is gone: In adult elite sports teams and in university admissions, for example, individuals born early in the year are often overrepresented (e.g., Addona & Yates, 2010; Bedard & Dhuey, 2006; Schorer et al., 2009). This is most successfully explained by the maturation-selection hypothesis (Cobley et al., 2009), which argues that the initial selection earlier in life provided increased opportunities and training. Other important theoretical and practical work has explored the complex effects arising from annual variation and different types of age group structures (Schorer et al., 2013).

At the time of this initial selection, however, and in general settings such as school classes, a very simple model of RAEs may be adequate. RAEs can be understood as the mathematically inevitable result of imposing age groups in a context where ability varies continuously with age. In a previous article, we considered RAEs in a general setting for one measure of cardiorespiratory fitness, the 20 metre shuttle run (Veldhuizen et al., forthcoming A), and argued that RAEs arose simply from the underlying relationship between age and ability, with no net effect for other mechanisms. We also reported, in general terms, on the apparent sizes of relative age effects for boys and girls of different ages, noting that, as existing literature suggested, they were largest for adolescent males.
In this article, we further explore RAEs, using the same example. The shuttle run is of specific interest in pediatric exercise and physical education (it is, for example, widely used in fitness assessments; Ernst et al., 2006), and is also used to assess maximal VO$_2$ in sport contexts (e.g., Till et al., 2011). Mostly, however, it is a convenient model. It supplies a relatively simple ability that increases with age, and which has some relevance to athletics, which is the area in which RAEs have been most widely-reported and studied. The shuttle run is also a simple fitness measure that is perhaps less likely than other kinds of performance to reflect specific training, coaching, access to resources, or teacher expectations. Growth curve modeling based on longitudinal data covering several years should also produce estimates of age-related variation that are, in any case, essentially independent of such factors.

In this analysis, we develop growth curve models (Singer & Willett, 2003) that allow us to calculate normal scores based on an individual’s exact age. We use these results to quantify RAEs in two different contexts. First, we calculate the advantages and disadvantages conferred by birth at different points in the year when individuals are ranked or graded relative to a norm. Second, we examine RAEs in situations involving the selection of individuals with high or low levels of ability. We argue that RAEs of the magnitude we report produce only minor errors in grading or ranking, but that the same RAEs can give rise to large age-related differences in the probability of being identified as exceptional.
Methods and analysis

Data

Data for this study come from the Physical Health Activity Study Team (PHAST) project, a large, longitudinal study of health, motor functioning and associated characteristics conducted between September 2004 and June 2009. Details on this study are available elsewhere (Cairney et al., 2012). Briefly, PHAST comprised an open cohort of 2278 children measured on 9 occasions between the ages of 9 and 15. Our data include 15,240 assessments. In this analysis, we focus on males only, which reduces the number of data points to 7732. Among females, we previously found no association between age and performance after about age 12; RAEs in this group would therefore be small.

Shuttle run

At each assessment, study participants completed a 20m shuttle run test. The shuttle run is a widely-used test of cardiorespiratory fitness (Leger & Lambert, 1982; Leger et al., 1988) that involves running back and forth between two lines set 20m apart in synchrony to a beep, with the time between beeps becoming progressively shorter until the participant cannot continue. The number of stages completed is a measure of cardiorespiratory fitness that can be combined with age and sex to estimate maximum volume of oxygen consumed (Léger et al., 1998; Melo et al., 2010; McCammon et al., 1997; Van Mechelen et al., 1986). Shuttle runs were conducted in school gymnasias. Data were collected twice a year, in spring and fall, for grades 4 through 8, and once in
grade 9. RAEs have previously been shown to exist for this measure (Roberts et al., 2012).

Analysis

We re-analyzed our data with the aim of developing a method for converting all performances to normal scores, which would allow us to place all individuals on a common metric and to directly calculate expected RAEs. This requires a means of estimating the expected value and the expected variance at any age.

To calculate the expected number of stages completed for each child (i.e., the predicted mean at the child’s exact age), we used the fixed effects of the growth curve result previously reported, which is $2.06 + 0.017\times\text{age}^2$ (Veldhuizen et al, forthcoming A). To estimate the variance, we used locally weighted regression (LOWESS) to obtain an expected mean at every age, and then subtracted this expected mean from the observed performances to produce deviation scores. LOWESS is a curve-fitting method that provides a highly smoothed, non-linear estimate of the relationship between age and performance without parametric assumptions. We used the resulting variances as the outcome of a mixed-effects model including fractional polynomial transformations of age. The resulting equation represents the association between age and variance. Solving this equation at a point of interest produces an expected mean deviation, which can be converted to a standard deviation (SD) by dividing it by the normalizing constant $(2/\pi)^{0.5}$. 
The two fixed-effects equations provide a way to calculate the expected value and the expected variance for any age. We used these equations to calculate relative age effects in various situations, by substituting in ages and ability levels, obtaining Z scores, and converting these to percentile ranks. Specifically, we calculated the expected levels of performance for children of various ages born at different points in the year, and then determined where that level of ability would fall on the distribution for children born at the mid-point of that year (which, given the nearly linear association, closely approximates the expected mean for a group of children with the same year of birth). The difference between the two percentile ranks represents the expected error when an assessment where individuals from a single birth year are compared to one another. We used Stata 13 for all analyses (StataCorp, 2013).

**Results**

The equation for estimating mean performance, as above, was $2.06 + 0.017 \times \text{age}^2$. Fixed effects from a second growth curve model for the absolute deviation yielded the equation (with age in decimal years): $\text{AD} = 0.0367 + 0.00644 \times \text{age}^3 - 0.00216 \times \text{age}^2 \times \ln(\text{age})$.

Model results are illustrated in Figure 1. Ability increases almost linearly with age, while the variance function is approximately linear at early ages, but flattens out among older ones. Model results produce expected means that rise from 3.4 completed shuttle run stages at age 9 to about 6.0 at age 15, while expected SDs rise from 1.8 to 2.7 over the same period, indicating greater variability in performance with increasing age. Examining crude means and SDs for various arbitrary age groups in our sample
confirmed that these estimates described the actual age-related changes well. These models imply an effect size for a one-year difference in age of about 0.20.

Figure 3-1. Predicted mean and variance for Leger shuttle run performance by age.

We used model results to calculate expected RAEs under a variety of scenarios. Table 1 shows the errors in percentile ranks for children of different ages and ability levels. Columns are shown for percentiles corresponding to -2, -1, 0, 1, and 2 SDs from the mean. Values are the difference between the true rank of each child and the ranking they will receive when compared to the norms for their year.
Table 3-1. Predicted errors in percentile ranks for the Leger shuttle run test, by exact age and actual ability level.

<table>
<thead>
<tr>
<th>Age in years</th>
<th>Exact age</th>
<th>Actual ability (percentile; SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2.3 (-2 SD)</td>
</tr>
<tr>
<td>9</td>
<td>9.01</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>9.25</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>9.75</td>
<td>-0.2</td>
</tr>
<tr>
<td></td>
<td>9.99</td>
<td>-0.4</td>
</tr>
<tr>
<td>10</td>
<td>10.01</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>10.25</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>10.75</td>
<td>-0.1</td>
</tr>
<tr>
<td></td>
<td>10.99</td>
<td>-0.2</td>
</tr>
<tr>
<td>11</td>
<td>11.01</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>11.25</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>11.75</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>11.99</td>
<td>0.0</td>
</tr>
<tr>
<td>12</td>
<td>12.01</td>
<td>-0.1</td>
</tr>
<tr>
<td></td>
<td>12.25</td>
<td>-0.1</td>
</tr>
<tr>
<td></td>
<td>12.75</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>12.99</td>
<td>0.2</td>
</tr>
<tr>
<td>13</td>
<td>13.01</td>
<td>-0.3</td>
</tr>
<tr>
<td></td>
<td>13.25</td>
<td>-0.2</td>
</tr>
<tr>
<td></td>
<td>13.75</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>13.99</td>
<td>0.4</td>
</tr>
<tr>
<td>14</td>
<td>14.01</td>
<td>-0.5</td>
</tr>
<tr>
<td></td>
<td>14.25</td>
<td>-0.3</td>
</tr>
<tr>
<td></td>
<td>14.75</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>14.99</td>
<td>0.8</td>
</tr>
</tbody>
</table>

As an example, we can consider the norms for children aged between 9 and 10 years. The mean can be closely approximated by simply considering expectations for a child at the median age, which is 9.5 years. This gives $2.06 + 0.017\times9.5^2 = 3.58$ stages. The SD at this age can be calculated by solving the equation above and dividing by the normalizing constant, which gives 1.73 stages.

We can then consider the case of a perfectly average child aged 9.01 years who is evaluated against this standard. His performance will be $2.06 + 0.017\times9.01^2 = 3.43$ stages. This, however, is 0.15 stages below the norm of 3.58. When comparing this child
to the norms for his age group, then, he appears to be 0.15/1.73 = 0.09 SDs below the mean. This corresponds to the 46.5th percentile. The grading error, therefore, is 3.5 percentile ranks: This average child appears very slightly below average because of his relative youth.

Discussion

We can distinguish between two contexts in which RAEs may occur. One concerns the grading or ranking of individuals within a general sample. The other concerns the selection of individuals with exceptionally high or low levels of performance.

RAEs and ranking or grading

When age groups include a maximum one-year difference, and the effect size of the age-ability relationship is comparable to that in our results, RAEs produce fairly small ranking errors. Differences between actual and apparent performance are never greater than about 4 percentile ranks. These errors are greatest wherever the distribution is densest. For a distribution with tails (i.e., one in which large or small values are less common than moderate ones), it is individuals close to the average who are most affected in grading or ranking contexts.

For other tasks or in other populations, of course, RAEs will vary with the strength of the underlying relationship between age and ability, and with other factors, such as previous identification as exceptional (which, in sports contexts, can provide greater opportunities for skill development; Helsen et al., 1998). Our results, however, are roughly consistent
with the magnitude of RAEs reported in general educational contexts (reviewed in Cobley et al., 2009b).

**RAEs in the identification of exceptional individuals**

Age differences, however, have been shown to be fairly large in groups that have been identified as exceptional. In a recent meta-analysis on elite sport, Cobley et al. (2009a) reported that people born in the first quarter of the year outnumbered those born in the last quarter by about 3:2. Larger effects have also been reported; the first article on RAES in sport noted that National Hockey League players born in the first quarter of the year outnumbered those in born in the fourth by almost two to one (Barnsley, et al., 1985). Similar results have been obtained in other settings. Morrow et al., for example, reported that boys born in December were 1.3 times, and girls 1.7 times, more likely than those born in January to have received a diagnosis of ADHD (Morrow et al., 2012). One question, then, is why RAES seem to be fairly small when individuals are ranked or graded, but large when we consider age gradients among people identified as exceptional.

One explanation is simply that the tasks or populations are different. The relationship between age and ability may be stronger in other settings or for other abilities. Alternatively, the relationship between age and absolute performance may be similar, but the variance lower. RAES arising from ordinary age-related variation are affected by both the slope of the age effect and the variance (with a linear association and a normal distribution, they are, in fact, a simple function of effect size). In our sample, fitness levels varied tremendously; a one-year increase in age was associated with a change of
about half a shuttle run stage, but the overall standard deviation was roughly 2 stages. In a sample with less variance – for example, one in which fitness was universally good – RAEs would be larger. In our results, a male child with average ability for his exact age would appear at about the 47th percentile of the class if born at the end of the year, and at about the 53rd if born at the beginning. If the SD were, instead, only 0.5, then these ranks would be the 34th and the 67th.

A possibly more useful way to consider the situation, however, concerns the difference between ranking and selection. Selecting people from the extremes of the distribution can give rise to large age differences even if the actual rankings are nearly accurate. Any cut point applied to make a selection will lie farther from the expected level of ability for younger than for older children. In most distributions likely to be observed in these contexts, the expected frequency drops off sharply at the extremes. This means that the same age differences that produce only small errors in grading can produce large relative risks of qualifying for “exceptional” status.

This can be most easily demonstrated with an example. For statistical convenience, we will pretend that shuttle run performance is normally distributed (in fact, it is only vaguely normal). To be consistent with existing articles on RAEs, in which children are often divided into groups according to the quarter of the year in which they were born, we can consider a group of 12 year old children: Some born midway through the first quarter of the year (aged 12.875), and some midway through the last (aged 12.125). All are candidates for an elite athletics team whose eccentric coach uses completion of stage 12
on the 20m shuttle run as the sole criterion for selection. The equations reported earlier let us determine that, for the younger children, stage 12 requires a performance 3.10 SDs above the level expected for their age. Only 0.096% of the normal distribution lies at or beyond 3.10 SDs. For the older children, however, stage 12 is “only” 2.86 SDs above the mean; and 0.21% of observations can be expected to be at this level or higher. On the team finally selected, therefore, we would expect the older children to outnumber the younger ones by the ratio 0.21/0.096, or about 2.1 to 1. This ratio is related directly to the difficulty threshold. If we required stage 8, for example, it would be only 1.4 to 1, and if we merely required performance above the median, it would be only 1.1 to 1. Whereas grading or ranking errors are largest where the distribution is densest, then, the age-related relative risks of selection are greatest at extreme values. Although the absolute proportion of individuals affected will be small, relative age differences will come to strongly affect the probability of being selected.

This phenomenon is illustrated in Figure 2, which shows hypothetical distributions for 3 groups, each representing children born at different points in the same year. At a cut-point of 11, older children will outnumber younger ones only slightly. This can be seen by looking to the right of 11 and comparing the area under the dotted line to the area under the solid line. The further to the right we place the cut-point, however, the greater the difference becomes.
Our examples, of course, reflect situations in which age and performance are related in the same way at all levels of performance, and in which the relevant trait is normally distributed through the extreme tails. Neither of these things are particularly likely to be true in real-world situations; but they do not need to be for the general principle to be valid.

This view may also be of some use in explaining RAEs in other contexts, for example, in the selection of students for gifted programs or – the effects at the left of the distribution being the same as those at the right – for diagnosis of mental health problems (Reijneveld
et al., 2006). Although developmental disorders or other syndromes may give rise to
different distributions in the left “tail”, the same general relationship may be present in
some form when criteria for developmental delay or associated conditions are applied
(see Morrow et al., 2012, for an example of RAES in this context). This is perhaps one
reason to prefer screening and assessment instruments that permit the calculation of
scores for exact child ages over those that supply norms for broad age ranges (see also
Veldhuizen et al., forthcoming B).

All this, of course, concerns only the initial selection of exceptional individuals. The
 persistence of RAES into adulthood requires separate consideration. The most
compelling interpretation is that, as existing work has argued, the initial selection
produces a ‘streaming’ of individuals that continues to determine their success (see Baker
et al., 2010, for a review). In elite sport, for example, being selected for a team in
adolescence is essentially a prerequisite for success at a professional level in adulthood;
by the time RAES are no longer operative, it will be too late for a talented individual born
late in the year to enter the system.

Limitations

Although analysis of the underlying age-ability association provides a simple and useful
way to examine these phenomena, this approach clearly cannot fully account for RAES in
all contexts. Most things in life are more complicated than running back and forth
between two lines, and there will often be scope for variation due to other factors. Our
results imply, though, that a perfectly straightforward association between age and ability
seems capable of giving rise to quite large effects whenever selections of this kind are made, and that other explanations may not always be necessary.

**Conclusion**

A small-to-moderate association between age and performance is unlikely to cause individuals’ abilities to be grossly over- or under-estimated when they are graded or ranked relative to peers. The *same* effects, however, can produce large differences in the probability of being identified as exceptional, and to relatively large age gradients within these groups of exceptional individuals. As a result, some talented individuals may be denied opportunities they merit. At the other extreme of the distribution, relatively young individuals may considerably more likely to be identified as having a disorder, and, conversely, relatively old ones may be denied assistance.
References


StataCorp (2013). Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.


CHAPTER 4: Misclassification due to age grouping in measures of child development.

Scott Veldhuizen, MA$^{1,2,3}$, Christine Rodriguez, MSc$^2$, Terrance J. Wade, PhD$^3$, John Cairney, PhD$^{2,3}$

Affiliations: $^1$Centre for Addictions and Mental Health, Toronto, Ontario, $^2$Department of Family Medicine & Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton Ontario, $^3$Department of Health Sciences, Brock University, St. Catharines, Ontario

Introduction

Developmental delay is thought to be present in perhaps 13% of children (Rosenberg et al., 2008), and screening for it in primary care, though still the subject of debate (American Academy of Pediatrics, 2001; Baker & Lecture, 2011; Hertzman et al., 2011), has become widespread. Although unstructured assessments can also be done, standardized instruments are widely used in both clinical and research settings. Such instruments commonly evaluate children on a set of items, convert raw scores to standard ones using age-specific norms, and then apply cut-points to determine caseness.

One problem with this process is that norms and cut-points are applied by age group: One threshold will be applied to all children aged 9 months to 12 months, another to those >12 months to 15 months, and so on. Development, however, is a continuous process, without sudden changes at specific ages. This leads to what are known in some contexts as relative age effects (RAEs): Because age groups have been imposed, the oldest children within each group can be expected to perform well, on average, while the youngest are disadvantaged. This is analogous to the RAEs known in other contexts, which have been shown to affect diagnosis of attention deficit hyperactivity disorder (Morrow et al., 2012), academic success (Cobley et al., 2009a), and participation in elite-level sport (Cobley et al., 2009b).

The use of age bands means that a child’s age may determine whether he or she meets criteria for caseness. This is illustrated in Figure 1. Children close to the threshold are misclassified, with the direction depending on their age within their band. Although
these misclassifications will usually affect marginal cases (a child with above-average development will not usually appear to be severely impaired, or vice versa), this may still lead to misreferrals, and to the misallocation of clinical resources. It can also be argued that the value of standardized screens lies in their ability to make exactly these types of distinctions; discriminating between normal development and severe delay does not necessarily require a standardized instrument.

Figure 4-1. Illustration of misclassifications resulting from relative age effects. (Note: FN = false negative; FP = false positive.)

Age banding can also worsen agreement when instruments that use different sets of age groups are compared. The cases identified by each will depend not only on the ability of each child and on the measurement properties of the instrument, but on each child’s
location within each of the two age groups being applied. This may limit the measured agreement between measures in studies of concurrent validity, and may lead to diagnostic confusion if multiple instruments are used clinically. This is illustrated in Figure 2, which shows disagreements between two otherwise identical measures that use different age groups. Children falling in the shaded areas will have different results on each measure.

Figure 4-2. Illustration of disagreement between instruments arising from use of different age groups. (Note: Children falling in the shaded areas will have different statuses on each instrument.)

Using age group-specific norms, then, offers simplicity at the cost of accuracy. Growth curves (of the kind used to evaluate children’s height and weight against typical values)
are one alternative to this approach. Computerized scoring, where practical, should make it possible to improve further on the accuracy of results.

Although age grouping clearly limits the accuracy of assessments of development, the actual rates of misclassification it produces have not, as far as we know, been measured. In this article, we estimate these error rates for the cognitive subscale of the third edition of the Bayley Scales of Infant and Toddler Development (BSID-III; Bayley, 2006). We use this measure as an illustrative example. Misclassification rates will vary across instruments, samples, and populations, but all measures that use age groups for norming will suffer from this effect to some extent.

**Methods**

**Data**

Data come from the Psychometric Assessment of the NDDS (PANS) study. PANS was conducted in 2012. Its purpose was to validate the NDDS-2011, a revision of the Nipissing District Developmental Screening tool (NDDS; NDDS, 2013), in a general population sample. A total of 813 children aged 1 month to 6 years were recruited from community organizations in the vicinity of Hamilton, Ontario, Canada, and assessed with a battery of measures, including the BSID-III. The BSID-III was administered by research assistants with undergraduate or graduate degrees, who completed a minimum of 8 hours of training and 10 hours of supervised test.
Children were excluded from PANS only if they had physical health impairments representing a serious barrier to testing (e.g., blindness, deafness). In practice, no referred children were excluded for this reason. Parents had to read and speak English and be the legal guardians of the child. Four children were excluded due to language barriers, and 4 because children became distressed and could not complete the appointment. We adjusted child ages for pre-term birth, as recommended in the BSID-III manual (Bayley, 2006). The BSID-III covers the age range from 1 month to 42 months, 15 days. In PANS, 593 children (305 male, 288 female) were in this range and received the BSID-III. These 593 subjects are included in the present analysis. Parents of PANS participants were slightly better-educated, and had slightly higher incomes, than the national average (Cairney et al., 2012; Veldhuizen et al., 2015). We received ethical approval from the McMaster University Research Ethics Board, and all parents provided informed, written consent.

**BSID-III**

The BSID-III is a clinician-completed instrument intended for infants and children aged between 16 days and 42 months, 15 days. It comprises 325 items divided into 5 subscales (cognitive, receptive communication, expressive communication, fine motor and gross motor). The cognitive subscale includes 90 items. Not all items are used for all children; the child’s age determines where in the list the assessment begins. Administration of the test ends after 5 consecutive items are missed. The number of items passed or skipped is the raw scale score. The age bands for the third edition of the
BSID-III are 10 days wide among young infants, rising to 1 month for those 5 months, 16 days or older, and to 3 months for children 3 years, 16 days or older (Bayley, 2006).

To explore the impact of using wider age bands on classification accuracy and on agreement between measures, we also applied the age groupings used by two parent-completed instruments: The Ages and Stages Questionnaire (ASQ) and the Nipissing District Development Screener (NDDS). Unlike the BSID-III, the ASQ and NDDS provide different sets of items for each age group (the fact that new versions must be developed for each group is presumably one reason they use broader age bands). Because all three measures apply the same cut-points to an entire range of ages, however, the underlying problem is the same for all.

The ASQ uses intervals of 2 months for infants, of 1 month for children 8 months to 10 months, of 3 months for those 2 years of age, and of 6 months for children 3 years and older (Squires et al., 2009). The NDDS uses 2 month intervals for infants, 3 months intervals for those 6 months and older, 6 month intervals for those 18 months and older, and 1 year intervals for those over 3 years of age (NDDS, 2013). Hereafter, we describe the ASQ bands as “moderate” and the NDDS bands as “wide”.

**Analytic approach**

Our basic approach is to use a real dataset of BSID-III cognitive subscale raw scores to model the expected value and variance of this measure as functions of age. Provided these functions fit the data well, this allows us to impose any set of age groups, estimate
‘norms’ and cut-points for each, and find the volumes of the ‘false positive’ and ‘false negative’ regions shown in Figure 1. We use simulation to do this, which helps to illustrate the situation while providing solutions through a kind of simple Monte Carlo numerical integration. Imposing different age groups then lets us estimate the level of disagreement between instruments that is likely to result from age grouping differences alone.

**Analysis**

We first regressed cognitive subscale score on child age using the 593 children from the PANS study. To capture the non-linearity of this association, we used the method of fractional polynomials (Royston et al., 1999). As the subscale scores showed signs of heteroscedasticity (see Figure 3; note that clustering of cases by age is the result of recruitment effects), becoming more spread out with increasing age, we also regressed the absolute values of the residuals from this model on age, again using fractional polynomials. This produced a second equation that provides an estimate of the expected mean deviation for a given age. We converted the mean deviation to a standard deviation by multiplying it by the normalizing constant \[1 / (2/\pi)^{0.5}\]. Together, these equations give estimates of the mean and variance for any child age.

We then generated a dataset containing one million simulated participants. To do this, we drew one million ‘ages’ from a uniform distribution covering the age range of the BSID-III. We calculated the expected mean and SD of the cognitive subscale score at each of these exact ages using the regression equations in Table 1. We then generated a
cognitive subscale score for each simulated participant by drawing a random number from a normal distribution centred on the expected mean and with variance given by the expected SD. We imposed the BSID-III age bands and calculated 1) the “norms” for each age group; and 2) the status of each simulated participant based on these norms. Finally, we counted the simulated participants whose true status and norms-based status disagreed.

To test the effects of the wider age bands used in other instruments, we then imposed the moderate and wide age groupings on the same dataset of simulated participants. To assess the sensitivity of validation studies to differences in age bands, we then compared classifications resulting from the BSID age bands to those produced by the moderate (ASQ) and wide (NDDS) age bands. This produces an estimate of the disagreement we might expect between two otherwise identical measures that simply use different sets of age groups. We performed the analysis using Stata 13 (StataCorp, 2013) and R 2.14 (R Development Team, 2008).

**Results**

One-term polynomial transformations of age produced good fits in both regression models (Table 1). Additional polynomial terms were significant for the cognitive score model, but we omitted them from the final model because the very minor improvements in fit they produced seemed not to justify the complexity ($R^2$ for one-term model = 0.973; for two-term model, 0.975). A histogram of residuals confirmed that they were roughly normally distributed, and that drawing random numbers from a normal distribution would
therefore provide a good approximation to the data in our sample. A random selection of 593 of these data points is shown in Figure 4. Comparison to the PANS data in Figure 3 confirms that this process produces plausible data.

Figure 4-3. BSID-III cognitive subscale raw scores by adjusted age from PANS study (n=593).
Figure 4-4. Example of generated data (n=593).

Table 4-1. Linear regression models for BSID-III cognitive subscale mean and variance.

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>Robust SE</th>
<th>t</th>
<th>P</th>
<th>L95%</th>
<th>U95%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Cognitive score model.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age(^{0.5})</td>
<td>14.75</td>
<td>0.10</td>
<td>144.27</td>
<td>&lt;0.001</td>
<td>14.55</td>
<td>14.95</td>
</tr>
<tr>
<td>Intercept</td>
<td>-10.46</td>
<td>0.39</td>
<td>-24.73</td>
<td>&lt;0.001</td>
<td>-11.23</td>
<td>-9.69</td>
</tr>
<tr>
<td><strong>B. Residuals model.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.0197</td>
<td>0.009</td>
<td>2.29</td>
<td>0.023</td>
<td>0.0028</td>
<td>0.0365</td>
</tr>
<tr>
<td>Intercept</td>
<td>2.72</td>
<td>0.16</td>
<td>17.17</td>
<td>&lt;0.001</td>
<td>2.40</td>
<td>3.03</td>
</tr>
</tbody>
</table>

Of our million randomly-generated data points, 2.3% (as expected) were below the -2SD cut-point. When we divided the sample into BSID-III age groups and calculated group-specific cut-points, 15% of the ‘cases’ were misclassified as negatives. Conversely, 15% of the cases that were below the age group cut-points were false positives. These results imply that, in a sample resembling ours, even a perfect measure of development would
have its sensitivity reduced to 85% due to the use of age groups. With moderate (ASQ) age bands, 27% of identified cases were false positives, and 28% of true cases were missed. With wide (NDDS) bands, 44% of identified cases were false positives, and 46% of true cases were missed.

To examine how different age groups might limit measured agreement, we then compared classifications based on the moderate (ASQ) bands to those based on the narrow (BSID-III) bands. This represents the agreement that would be observed if the measures were identical and free from random error, and should approximate the limitation on agreement attributable simply to the use of different age groupings. We obtained a pseudo-sensitivity of 69%, and a pseudo-specificity of 99%. Repeating this process for the wide (NDDS) age groups yielded results of 53% and 99%, respectively, relative to a reference measure using the narrow (BSID-III) bands.

**Discussion**

Results show that age banding is an important source of error in assessments of development. Even in the BSID-III, which provides norms for quite narrow age bands, 15% of true cases are likely to be missed as a result of grouping children by age. As we have noted, these are generally the milder cases of delay, while the false positives will be among the most delayed non-cases. These misclassifications are therefore less serious clinically than they would be if they represented errors occurring completely at random. Any misclassification, however, will lead to sub-optimal use of resources. Stigma may
also be a concern for false positives, while false negatives could lead to a potentially beneficial intervention not being offered.

Instruments using wider age bands have higher misclassification rates. In our data, the moderate (ASQ) age bands resulted in 28% of cases being missed, while the very wide NDDS bands caused almost half to go undetected. This is not altogether surprising, considering the wide range of ages to which the same norms are sometimes applied. On the ASQ and NDDS, a single cut-point is used for children aged 9 to 12 months. In our models, a child aged 9 months would be expected to score about 2 SD below one aged 12 months (or 1 SD below the age group mean). Misclassifications for wider age bands can also be expected to be more serious, with considerably larger errors possible for individual children.

Our results also suggest that differences in age groups may be an important source of disagreement between measures in validation studies and in clinical settings. This is a consideration, also, for the development of new measures, which may be limited in their ability to show acceptable agreement with existing ones. In principle, instruments with similar age groups may agree well, despite substantial errors relative to children’s true status, because they will misclassify the same children; in our examples, however, differences between age groups produced high levels of disagreement. We can consider, for example, the hypothetical situation of two otherwise identical instruments, one with the age bands of the ASQ, and one with the age bands of the BSID-III. In a sample of
children covering the full age range of the latter instrument, the measured sensitivity of the former relative to the latter is likely to be limited to about 70%.

Although there are considerable differences between our examination and real validation studies (notably, we considered only a single subscale), age banding may also help to explain the often modest agreement reported in comparisons of the BSID-III and ASQ (Simard et al., 2012; Gollenberg, 2006; Kim & Sung, 2007; Woodward et al., 2011) and of the BSID-III and NDDS (Cairney et al., 2012; Veldhuizen et al., 2015; Dahinten & Ford, 2004). Although sensitivities above 70% have been reported in comparisons of the ASQ and BSID-III, these are often from studies that do not include broad age ranges. Significantly, the ASQ’s published -2 SD thresholds also almost invariably identify far more children than the expected 2.3%, even in general population samples (e.g., Limbos & Joyce, 2011; Squires et al., 1997). This will lead to higher measured sensitivity (at the expense of lower specificity).

It is important to note that instrument developers are aware of the problems posed by age banding, and that this issue is just one of many difficulties that arise when constructing standardized measures of development. Historically, providing norms and scoring guidelines for age groups has had important advantages: It makes paper-and-pencil scoring practical, and it is natural when different versions of the instrument have been developed for different age ranges. Scoring and data entry are now increasingly computerized, however, and, in this context, deriving scores that are fully adjusted for
child age is much less difficult. When computerized scoring is not practical, growth
curve charts (Wei et al., 2006) offer another alternative.

Limitations

We considered only one developmental domain and one measure. Effects will be
somewhat different with other measures and in other samples. Although our approach fit
our data well, residual distributions may be non-normal in other settings, particularly if
there are many cases of relatively severe delay present. We also applied age bands from
two other instruments to an entirely different measure; this allowed us to explore the
effects of wider age groups, but, because these instruments will vary in other ways,
results are only illustrative.

Conclusion

The use of age groupings in standardized tests of development leads to considerable and
avoidable errors in classification. Computerization makes it practical to promote scoring
systems that avoid the use of age groups.
References


StataCorp. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP, 2013.


CHAPTER 5: Errors, misclassifications, and instability due to the use of age bands in measures of child development

Scott Veldhuizen\textsuperscript{1}, MA
John Cairney\textsuperscript{1,2}, PhD
Terrance J. Wade\textsuperscript{3}, PhD

Affiliations:
\textsuperscript{1} Department of Family Medicine, McMaster University, Hamilton, Ontario, Canada
\textsuperscript{2} Departments of Psychiatry and Behavioral Neurosciences and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
\textsuperscript{3} Department of Health Sciences, Brock University, St Catharines, Ontario, Canada
Introduction

Some 12-16% of children have delays in cognition, language, or motor functioning (Rosenberg, 2008). These children can be identified in routine clinical care, but many standardized measures also exist for this purpose, and for the monitoring of development generally (Grant et al., 2010). Although some instruments consist only of expressions of parental concern (PEDS; Glascoe, 1997), many of those in widespread use are essentially checklists of milestones or behaviours, sometimes separated explicitly into domains (usually including cognitive, motor, and communication). A count of missed milestones may be used to indicate a problem, or may be translated into a standard score or a ‘developmental age’.

These instruments, which include widely-used screens like the Ages and Stages Questionnaires (ASQ; Squires, 2009) and clinician-completed reference instruments like the Bayley’s Scales of Infant and Toddler Development (BSID; Bayley, 2006), are usually scored by sorting children into age groups and (essentially) applying a single cut-point to all children in each group. (Exceptions do exist, such as the Child Development Inventory, which converts raw scores directly to ‘developmental ages’; Ireton, 1992). This approach to scoring leads to systematic error: The level of development will be underestimated for the younger children within each age group and overestimated for the older ones. It has long been appreciated that some children ageing into a new scoring band may suddenly become cases (Boyd, 1989), but this is only part of the problem; age banded scoring will cause some degree of error for essentially everyone who is assessed.
In a previous article (Veldhuizen et al., 2015a), we estimated misclassification rates resulting from age banding using as an example a subscale from the BSID-III. We also noted that the use of different sets of age groups will limit the possible agreement between instruments, and suggested that this may be one reason for the modest results reported by validation studies (e.g., Dahinten & Ford, 2004; Limbos & Joyce, 2011).

This article extends that work in three ways. First, we examine the performance of measures that include multiple developmental domains, again using the BSID-III as an example. Second, we consider the sizes of the errors involved: For children close to a threshold, even a small age banding error can produce a ‘misclassification’, but the consequences are arguably not as serious as they would be if a child with a severe delay were missed or one with no impairments were wrongly identified as delayed. Finally, we measure the effects of age grouping on longitudinal stability. A child whose level of functioning relative to immediate peers does not change can expect his or her scores on an age-banded measure to rise with age while they are within a band, and then fall when they transition to the next. This is likely to limit the agreement among assessments performed at different times, which may give the appearance of developmental discontinuity, and which could also lead to poor measured predictive validity of the instrument (e.g., Hack et al., 2005, Aylward, 2004).

**Methods**

The analysis consists of a simulation informed by real data. Our data come from the Psychometric Assessment of the Nipissing District Developmental Screen (PANS) study,
which includes BSID-III assessments conducted on 593 children (305 male, 288 female) aged 2 weeks to 6 years. Data were collected in 2012. Further details on PANS are available in other publications (Veldhuizen et al., 2015b).

We consider a child to be a ‘case’ of developmental delay if a score on any subscale falls below -2 standard deviations (SDs). As in our previous paper, we consider three sets of age bands: Those from the BSID-III itself, which we call ‘narrow’; those from the ASQ (‘medium’); and those from the NDDS (NDDS, 2013; ‘wide’). Age bands for the BSID-III are 10 days wide for infants, become 1 month wide for those over 5 months, 16 days, and 3 months wide for children over 3 years, 16 days. ASQ bands vary from 2 months to 6 months wide, and NDDS bands from 2 months to a full year (for children over 3).

Results for narrow age bands can be expected to approximate true BSID errors for samples resembling ours. We impose wider bands for the purpose of illustration; the fact that ASQ and NDDS results will not vary with age in exactly the same way as BSID raw scores, however, means that these results cannot be expected to closely resemble the actual errors affecting the ASQ and NDDS.

Our general approach is to use an initial analysis of the PANS data to determine how raw BSID-III domain scores change with age, and then use these results to construct a much larger simulated dataset. We then apply various scoring rules and age bands to measure quantities we would not be able to measure in the original sample. This simulation requires, at minimum, descriptions of the mean and variance of each domain, as well as
the correlations among domains, all as functions of age. We therefore perform the following analysis:

1. Regress age on the BSID raw score for each domain in turn, using the method of fractional polynomials with one and two terms to model non-linear associations and accepting the best-fitting model that is monotonic across the ages of interest.
2. Regress the absolute values of the residuals from (1) on age to obtain expected mean deviations as functions of age.
3. Transform these mean deviations into estimated SDs by dividing by the constant $1/(2/\pi)^{0.5}$ (Geary, 1935). This process produces a formula that can provide an expected SD for any age.
4. Convert raw domain scores into z-scores by subtracting raw scores from the predictions obtained in (1) and dividing by the SDs obtained in (3).

This part of the analysis produces functions that produce expected means and SDs as functions of age for each BSID-III domain, and converts the observed raw scores into z-scores. In the text stage, we determine how the relationships among the 5 domains vary with child age. To do this, we:

4. Divide the PANS sample into age deciles.
5. Calculate Pearson correlations for each pair of domains for each decile.
6. Regress the obtained correlation coefficients on age. (This is a basic approach that nevertheless should capture the general pattern. Alternatives exist (e.g., Altman, 1993),
but become difficult in the presence of negative correlations, which exist in our data for some domain pairs at early ages.)

We then create a large, synthetic dataset, as follows:

7. Create 785 simulated cases for each age (in days) between 16 (0.5 months) and 1290 (42.4 months) days, for a total of 1,000,875 simulated individuals.

8. Generate multivariate normal data for 5 domains for each one-day age group, with the age-specific correlation structure determined by the equations from (6). These generated data take the form of Z scores, and represent the “true” levels of functioning for each simulated individual.

9. Use the mean and variance equations from (1) and (2) to transform each domain Z score into a raw score.

10. Impose the 3 sets of age bands on the data. For each band, calculate a mean and SD from the raw scores from (9), and use these values to produce (a) a -2 SD threshold for each domain and (b) a post-banding standard score.

11. Compare scores and caseness after bands are applied to the ‘true’ generated scores.

Finally, we conduct a final analysis to examine longitudinal instability due to age banding:
12. Add 9 months to the age of each individual, repeat steps 9 and 10, and compare post-banding standard scores for each “time point”. (We use 9 month intervals as an essentially arbitrary example to illustrate the effect.)

Results

Correlations among domains were generally low among infants, and most rose more or less linearly through age 42 months (Figure 1). The exception was the association between cognitive and fine motor functioning, which was approximately 0.40 at all ages. As caseness is determined in our study by taking the lowest score on any domain (Bayley et al., 2006), the fact that correlations increase with age means that prevalence will decline with age. In our synthetic dataset, it is 13% in the youngest narrow (BSID-III) age band and 8% in the oldest.

Dividing records using narrow (i.e., BSID) age bands and calculating band-specific norms resulted in 10.4% of cases being misclassified as non-cases (i.e., false negatives) and 9.8% of apparent positives being false positives. For medium (i.e., ASQ) age bands, these numbers were 19.7% and 17.5%, and, for wide (i.e., NDDS) bands, 33.2% and 28.3%. These proportions are slightly lower than in the single-domain example we considered previously (Veldhuizen et al., 2015a).
As would be expected, the youngest children within each band were often false positives, and the oldest children false negatives. Figure 2 shows the proportion of cases misidentified as non-cases by exact age (in days) for narrow age bands. Although this proportion is 9.8% overall, it exceeds 60% for individuals close to the upper limits of certain age bands. For false positives, proportions are similar, but are obviously concentrated at the low end of age bands.

To assess the size of the measurement errors created by age bands, we averaged across the 5 domains and across all individuals and calculated mean absolute errors in SD units. The average error was 0.10 SDs for narrow bands, 0.19 SDs for medium bands, and 0.33
SDs for wide bands. This means that age banding tends, on average, to produce small but non-ignorable errors.

To measure the importance of errors among misclassified individuals, we calculated the mean error on the lowest single domain, as this is how cases were identified. For false positives, the lowest domain was underestimated by 0.19 SDs (maximum, 0.38 SDs) for narrow bands, by 0.38 SDs (maximum, 1.41 SDs) for medium bands, and by 0.58 SDs (maximum, 1.78 SDs) for wide bands. For false negatives, functioning was overestimated on average by 0.18 SDs (maximum, 0.58 SDs), 0.36 SDs (maximum, 1.52 SDs), and 0.54 SDs (maximum, 1.73 SDs), respectively. As would be expected, then, wider bands do not just increase the number of children misclassified; they also increase the sizes of the errors involved.

To illustrate the problem of longitudinal instability, we imagined a ‘retest’ at an interval of 9 months under the assumption that a child’s true percentile rank remained constant and the correlation among domains did not change. In this analysis, narrow bands led to 8.2% of cases becoming non-cases. Of the apparent cases, meanwhile, 7.4% had been non-cases at the previous time point. For medium bands, these proportions were 31.4% and 31.3%, and, for wide bands, 45.4% and 42.9%. This suggests that age banding alone can produce considerable instability in caseness (as Boyd (1989) has noted).
Another illustration of longitudinal instability can be obtained by calculating variation with age for a single individual. In Figure 3, we plot the percentile ranks that would be calculated for each set of age bands at each possible age for the BSID-III ‘cognitive’ subscale for a child whose true percentile is 50 at all times. This shows that an individual’s apparent functioning drops sharply as each new age band begins and then gradually rises. This illustrates how a child’s score can vary widely depending on his or her exact age. In percentile terms, these oscillations will be largest for a child who, like this one, is exactly average; they become smaller further away from the median. Many children with marginal functioning, however, would be expected to bounce back and forth across a threshold of ‘delay’.
Figure 5-3. Variation in calculated percentile ranks due to age banding for a child who is average at all ages (‘cognitive’ BSID-III subscale).

**Discussion**

In this study, we used model-informed simulation to explore measurement errors resulting from the use of age bands in measures of child development. In earlier work, we showed that age bands produce misclassifications; here, we extend this work to a multiple-domain measure, and also report on average errors for all children and on longitudinal instability.

In interpreting these results, it is important to bear in mind that we deliberately excluded other sources of measurement error. We also drew our simulated data from normal distributions, which means assuming that children with delay in a given domain are,
essentially, the left tail of a normal distribution. In some settings, there may be children whose development is affected by specific conditions such as Down syndrome. In this case, the appropriate distribution might be a mixture distribution, which could lead to a greater separation between cases and non-cases and, probably, a lower misclassification rate. Normal distributions are not unrealistic for general settings, however, especially as children with specific developmental syndromes might be expected to be identified without use of an instrument like the BSID-III.

**Age variation in correlations**

In our sample data, domain correlations generally increased with age. This is somewhat at odds with the correlations reported in the BSID-III technical manual, which show several U-shaped associations. However, estimates in the manual may have been inflated by age-related variation in functioning – the fact that domain scores vary with age means that the correlations among them will be increased when children of different ages are included (Goldstein, 1981). Our approach allowed us to calculate 'instantaneous' correlations.

One possible reason for rising inter-domain correlations is that these domains become better-integrated with age. Another is that measurement error is higher in younger children, perhaps because these domains are not well-developed or because they are more difficult to measure. Either way, the result is age-related variation in prevalence. When using the single lowest domain score to determine ‘caseness’, prevalence will tend to fall with age, and ‘cases’ tend to have more domains affected.
One interesting consequence is that, irrespective of age banding or other errors, some ‘cases’ will necessarily become non-cases with time. This measurement-induced decrease in prevalence may be of some concern, because many developmental issues (e.g., learning delays) either manifest or become easier to identify in older children. Identifying more children at earlier than at later ages, therefore, may be the opposite of what is desirable. However, several other diagnostic guidelines exist (those used in different jurisdictions in the United States are reviewed in Shackleford, 2006), and prevalence variation will differ for each.

**Misclassification rates**

Misclassification rates for narrow (BSID) age bands were somewhat smaller than those we reported previously for a single domain (Veldhuizen et al., 2015a). Misclassification rates were about 10% in each direction for narrow bands, about twice as high for medium bands, and three times as high for wide bands. The children most likely to be affected are those close to the caseness threshold, those close to an age-band limit, and those at ages where the change in scale scores is rapid. The latter two effects are illustrated in Figures 2 and 3.

It has been pointed out that the importance of misclassifications depends on the size of the errors involved (Glascoe, 2001). In our results, misclassifications under narrow bands were mostly caused by small errors (on average, about 0.2 SDs), while those for medium and wide bands were considerably larger, and are commensurately more important. A
child whose measured level of functioning is off by a full SD, for example, is clearly more likely to receive an inappropriate level of services (or to inappropriately receive no services at all). Any avoidable error is clearly problematic, however, because services will not be optimally distributed; even if a child at -1.9 SDs receives services while one at -2.1 SDs does not, this is also not ideal.

We also considered average errors for children who were not misclassified. These were generally small for narrow bands (0.1 SDs, on average), but more meaningful for medium (0.19 SDs) and wide (0.33 SDs) bands. The latter, for example, would move a child from the 50th percentile to the 37th, or from the 10th percentile to the 5th.

**Repeated assessments**

Age bands can create an illusion of instability over time. As Figure 3 illustrates, even narrow bands may cause development to appear to vary, depending on the exact timing of assessments. Particularly for wider age bands, some children can be expected to bounce back and forth across a clinical threshold. This variation may also be of some interest in the context of research that uses standardized instruments to look at developmental discontinuities. As age-banding errors will occur alongside regular measurement error, consistent results from repeated developmental assessments may be difficult to achieve. This could conceivably be one reason for the relatively poor predictive validity reported for developmental measures like the BSID (Hack et al., 2005).
Limitations and conclusions

We relied on a single sample for domain correlations and age effects, we assumed multivariate normality in our simulation dataset, we explored only 3 possible sets of age bands, and we did not attempt to incorporate other sources of error. Though we believe our results to be meaningful, they are illustrative rather than precise and generalizable. The real implications of age banding errors must also be judged in this context of imperfect measurement; if accuracy is poor for other reasons, age banding errors will be less important. Finally, our results on misclassification assume the blind application of a threshold; thoughtful clinicians may take exact child age into account in their judgements.

Implications

All age bands create systematic error, including false positives, false negatives, and apparent instability over time. Crucially, they are avoidable; they can be reduced by using narrower bands, or perhaps eliminated by moving to an approach based on growth charts (e.g., using the LMS approach (Cole, 1989)) or related methods.
References


10.1177/105381518901300202


Geary, R.C. The ratio of the mean deviation to the standard deviation as a test of normality. Biometrika 1935;27:310–32.


CHAPTER 6: Concurrent validity of the Ages and Stages Questionnaires and Bayley developmental scales in a general population sample

Scott Veldhuizen\textsuperscript{1,2}, MA, Jean Clinton\textsuperscript{3,4}, MD, Christine Rodriguez\textsuperscript{1}, MSc, Terrance J. Wade\textsuperscript{5}, PhD, John Cairney\textsuperscript{1,3,4}, PhD

Affiliations:
\textsuperscript{1} Department of Family Medicine, McMaster University, Hamilton, Ontario, Canada
\textsuperscript{2} Centre for Addiction and Mental Health, Health Services and Health Equity Research Department, Toronto, Ontario, Canada
\textsuperscript{3} Offord Centre for Child Studies, McMaster University, Hamilton, Ontario, Canada
\textsuperscript{4} Departments of Psychiatry and Behavioral Neurosciences and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
\textsuperscript{5} Department of Health Sciences, Brock University, St Catharines, Ontario, Canada

Introduction

One to three percent of children have a severe developmental delay such as a genetic syndrome or idiopathic autism or intellectual disability. A further 5 to 15% have milder but significant delays in learning, motor functioning, or other areas (Shevell et al., 2003; American Academy of Pediatrics, 1997). Early intervention can significantly benefit affected children (Blauw-Hospers et al., 2007; Anderson et al., 2010), but many conditions – particularly milder and more common problems, such as learning delays – are often not identified until children enter school (Kershaw et al., 2010). As such, there is considerable interest in systematic screening of young children for developmental delay. Routine screening has been recommended by the American Academy of Pediatrics (American Academy of Pediatrics, 2001), and has been implemented in some jurisdictions (e.g., the Ontario 18 month well-baby visit; Hertzman, 2008).

Screening can be done by clinicians, but clinical assessments by pediatricians and primary care physicians are relatively time-consuming and expensive (Dobrez et al., 2001). One alternative is to collect information from parents using standardized instruments. Parents are present during clinical visits, observe behaviors that may not be manifest in a brief clinical encounter, and do not bill health care systems for their time. Moreover, the information they supply has been shown to be useful (Glascoe, 2001), and most parent-completed instruments are brief, resulting in only a minor burden.

The Ages and Stages Questionnaires (ASQ; Squires et al., 1998) comprise one of the most popular such instruments. The current (third) edition consists of 21 different age-
specific, parent-completed questionnaires, each assessing child functioning in 5 domains: Communication, gross motor, fine motor, problem solving, and personal-social.

Published norms permit the identification of children falling various distances below the expected mean in each area. Children scoring below the -2 standard deviation (SD) cut-point on one or more domains are commonly identified as probable cases. Reports on the concurrent validity of the ASQ, however, have varied widely. Probably not coincidentally, studies on this question have differed considerably with respect to the ASQ edition, child age range, country, response rate, language, reference measure, sample size, and risk level of children.

Most validation studies have been limited by small sample sizes (e.g., Gollenberg et al., 2009; Squires et al., 2009), or included only children considered high-risk (e.g., Woodward et al., 2011; Lindsay et al., 2008; Schonhaut et al., 2013). Questions around the usefulness of the instrument in broad screening programs, however, are best answered by large, general population samples. Among studies on the English-language ASQ, such samples include those analyzed in Limbos & Joyce (2011) and in Hix-Small et al. (2007). Limbos & Joyce, reporting on a sample of 334 children assessed with a battery of clinician-administered instruments, reported a sensitivity of 82% and a specificity of 78%. The study of Hix-Small et al., though not a validation study, published results showing agreement between the ASQ and clinician agreement; these correspond to a sensitivity of 66% and specificity of 84%. The largest validation study of any kind was conducted by the instrument’s developers (Squires et al., 1997), who reported a sensitivity of 75% and a specificity of 86% in a mostly high-risk sample.
In the present study, we examine the concurrent validity of the ASQ with respect to the third edition of the Bayley Scales of Infant Development (BSID-III; Bayley, 2006) in a large, general-population sample of infants and toddlers. The BSID covers a set of domains similar to those evaluated by the ASQ, and is the reference measure most commonly used in studies on the validity of the ASQ (reflecting a scholarly consensus on its appropriateness). As the purpose of screening is usually to identify children who would benefit from further assessment or intervention, our primary analysis compares the overall status on both instruments: Children scoring below a referral cut-point on any of the BSID-III domains are considered cases, and children scoring below the -2 SD threshold on any ASQ domains are considered positive on the screen (i.e., potential cases of delay). Performance in specific domains is also of interest, and so we also explore the agreement of the problem solving, communication, gross motor, and fine motor subscales with measures of the same domains on the BSID-III (the remaining ASQ subscale, “personal-social”, has no clear counterpart on the BSID-III).

Methods

Study design and sample

Data are drawn from the Psychometric Assessment of the NDDS-2011 Study (PANS), a project concerned with the evaluation of another instrument, the Nipissing District Development Screener (NDDS, 2000). We recruited a convenience population sample from community organizations providing services to families in Hamilton, Ontario, and surrounding areas, between May 2010 and October 2011. Of 812 total participants, 218
were outside the age range of the BSID-III, and are excluded from our analysis for this reason. Of the remaining 594, 7 were excluded because the ASQ was completed after the BSID-III (n=2) or because appointment rescheduling led to a delay of more than 14 days between the two measures (n=5). The remaining 587 children form the sample in the present analysis.

Parents were eligible if they could speak and read English, and were the child’s primary caregiver and legal guardian, while children were excluded only if they had sensory impairments (e.g., blindness, deafness) or known genetic syndromes that would have prevented assessment. In practice, this exclusion affected no referred children. Parents received study materials in the mail. These consisted of the ASQ, the NDDS, a demographic survey, and, for parents of children aged 18 months or more, the Child Behavior Checklist (Achenbach & Rescorla, 2000). Parents were asked to complete all measures before their clinical appointment, and could do so in any order. In 36 (6%) cases, the ASQ completion date was missing (but is known from mailing dates to have been shortly before the clinical appointment). Of the remainder, 318 (58%) were completed on the same day as the BSID-III or on the day before, and 498 (90%) were completed within 7 days.

**Study design**

Child age was adjusted for prematurity if the child was under 2 years and born 4 weeks or more prematurely, and this adjusted age was used to determine the ASQ and BSID-III versions used. The BSID-III was administered by research assistants who were blind to
the results of the ASQ. Research assistants held undergraduate or master’s degrees in health sciences or social sciences, and received a minimum of 8 hours of training from an experienced pediatric psychometrist as well as 10 hours of supervised test administration. We received ethical approval from the McMaster University Research Ethics Board, and all parents provided informed, written consent.

As we noted in a related report (Cairney et al., 2016), the prevalence of delay in our sample, according to published BSID-III norms, was lower than anticipated and fell with increasing child age. All children, however, were recruited from the same sources and in the same way, and we did not see similar patterns on the ASQ. Significantly, concerns have been raised previously over low prevalences produced by the published BSID-III norms (Anderson et al., 2003). As these norms were developed with an enriched sample (Bayley, 2006), it is possible that they are problematic for a general population sample.

We therefore produced two sets of outcomes based on the BSID-III: One based on published norms and age groups, and one based on the observed distribution of raw scores. If the published norms are inappropriate, then a distribution-based measure should provide a better indicator of caseness. If the prevalence of delay genuinely falls among older children, however, then the published norms will better reflect their true status. If the concurrent validity of the ASQ differs between these two versions of the criterion measure, then more detailed analysis will be required.
Ages and Stages Questionnaires

For children up to 42.5 months, 17 different ASQ versions are relevant. Each includes a set of 6 items for each of the 5 domains. Items are scored as 0 (“not yet”), 5 (“sometimes”), or 10 (“yes”), and responses are summed to produce a score between 0 and 60. We used published norms for each version to identify children scoring 2 SDs or more below the expected mean and considered children positive on the ASQ if they scored below this threshold in one or more areas. We aggregated cases and non-cases across all versions of the ASQ before evaluating its agreement with our reference measure.

Bayley Scales of Infant and Toddler Development, Third Edition

The BSID-III includes 325 items divided into 5 domains: Cognition, receptive communication, expressive communication, fine motor, and gross motor. Items are approximately ordered by difficulty, and a child’s age determines the first item administered. The test continues until 5 consecutive items are missed. The number of items skipped or passed in each domain comprises the raw score, which is converted to a standard score using a set of published tables. We identified as cases those children who scored below the -2 SD (“extremely low”) cut-point on one or more domains.

Analysis

We performed two comparisons of the ASQ and BSID-III: One using the published BSID-III cut points, and one using cut-points based on the distribution of our sample.
To produce distribution-based thresholds, we used quantile regression with fractional polynomials. Quantile regression estimates a specified quantile as a function of the independent variables; whereas linear regression is concerned with the expected mean, quantile regression can estimate the expected median, or any other percentile (Yu et al., 2003). Entering age as an independent variable and specifying the 2.275\textsuperscript{th} percentile (which corresponds to 2 standard deviations below the mean of a normal distribution) makes it possible to estimate this threshold at the exact age of each child. As the relationship between age and BSID-III raw scores is not generally linear, we fit curves using the method of fractional polynomials (Royston, 2012). For BSID-III subscales, we used models with 1 or 2 age terms, subject to a condition of basic monotonicity (verified by examining plotted results). We performed this analysis for each of the 5 subscales.

We then calculated two indicators of delay: One identifying children falling below the published -2 SD cut-point on one or more BSID-III subscales, and one identifying those falling below the 2.275\textsuperscript{th} percentile on one or more subscales according to our distribution-based thresholds.

Finally, we produced 2x2 contingency tables comparing status on the ASQ to status on each of these BSID-III-based indicators, and calculated standard measures of agreement: sensitivity, specificity, and positive and negative predictive values. We performed this analysis for the overall result, as well as for four individual domains. We compared the ASQ fine motor and gross motor subscales to the BSID-III scales of the same names; “problem solving” to the BSID-III cognitive subscale; and “communication” to the
BSID-III “language” summary score produced by combining the “expressive communication” and “receptive communication” subscales. We calculated exact binomial confidence intervals for each measure of agreement, and tested the significance of agreement with Fisher exact tests. To examine agreement of the measures before application of cut-points, we computed Spearman correlation coefficients for each pair of subscales. We used Stata 13 (StataCorp, 2013) for all analyses.

Results

Most (92%) of our sample of "people most knowledgeable" were currently married or living common law, 80% had completed post-secondary education, and 50% had household incomes above $90,000. The national median household income in 2010 was $72,240 (Statistics Canada, 2014a), while the proportion of women aged 25-44 with completed post-secondary education was 73% (Statistics Canada, 2014b). These results imply that our sample had slightly higher socioeconomic status than is typical of the wider Canadian population. Five percent of children were reported to have been born 30 days or more before the due date. Complete descriptives are reported in Table 1.

Prevalence of developmental delay

Overall, 107 of 587 (18.2%) of children were positive on at least one ASQ subscale. There were 26 (4.4%) positive screens on the ASQ communication subscale, 28 (4.8%) on fine motor, 50 (8.5%) on gross motor, 36 (6.1%) on problem solving, and 33 (5.6%) on personal-social. These proportions are all significantly higher than the 2.3% expected at the putative threshold of -2 SD (all p<0.01). Prevalence on the ASQ did not vary
systematically with age, but was unexpectedly high among young infants: On the 2
month ASQ, 27 of 54 (50%) fell below the published cut-point on one or more subscales.

Table 6-1. PANS sample descriptives (total n=587).

<table>
<thead>
<tr>
<th></th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child's gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>305 (52%)</td>
</tr>
<tr>
<td>Female</td>
<td>282 (48%)</td>
</tr>
<tr>
<td>Prematurity</td>
<td></td>
</tr>
<tr>
<td>Within 30d of due date</td>
<td>558 (95%)</td>
</tr>
<tr>
<td>30d to 60d premature</td>
<td>23 (4%)</td>
</tr>
<tr>
<td>&gt;60d premature</td>
<td>6 (1%)</td>
</tr>
<tr>
<td>PMK Gender</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>579 (99%)</td>
</tr>
<tr>
<td>Male</td>
<td>8 (1%)</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
</tr>
<tr>
<td>Never married</td>
<td>31 (5%)</td>
</tr>
<tr>
<td>Married, common-law, or living with a partner</td>
<td>539 (92%)</td>
</tr>
<tr>
<td>Separated or divorced</td>
<td>14 (2%)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>Some secondary or less</td>
<td>25 (4%)</td>
</tr>
<tr>
<td>Completed high school or GED</td>
<td>27 (5%)</td>
</tr>
<tr>
<td>Some college or technical training</td>
<td>25 (4%)</td>
</tr>
<tr>
<td>Completed college or technical training</td>
<td>128 (22%)</td>
</tr>
<tr>
<td>Some university</td>
<td>40 (7%)</td>
</tr>
<tr>
<td>Completed a bachelor’s degree (BA, BSc, etc.)</td>
<td>208 (35%)</td>
</tr>
<tr>
<td>Completed a graduate or professional degree (MSc, MD, etc.)</td>
<td>134 (23%)</td>
</tr>
<tr>
<td>Household income (2009)</td>
<td></td>
</tr>
<tr>
<td>Under $35,000</td>
<td>75 (14%)</td>
</tr>
<tr>
<td>$35,000 to $59,999</td>
<td>73 (14%)</td>
</tr>
<tr>
<td>$60,000 to $89,999</td>
<td>114 (21%)</td>
</tr>
<tr>
<td>$90,000 to $129,999</td>
<td>161 (30%)</td>
</tr>
<tr>
<td>$130,000 or higher</td>
<td>116 (22%)</td>
</tr>
</tbody>
</table>
Published norms on the BSID-III, in contrast, produced prevalences that were lower than expected, ranging from 0.17% (1 case out of 587 children) on the “expressive communication” scale to 1.0% (6 cases) on the “gross motor” scale. These, too, were all significantly different from the expected 2.3% (all p<0.05). Overall, 17 of 587 children (2.9%) fell below the published “extremely low” (-2 SD) cut-point on one or more domains. As noted, prevalence also decreased with child age (overall point-biserial correlation = -0.09, p=0.03), with 15 cases out of 337 children under 18 months (4.5%) and only 2 cases (0.8%) of 250 older children.

The development of distribution-based thresholds is illustrated for one BSID-III subscale in Figure 1. This approach resulted in higher prevalences, with 12 or 13 children positive (2.0% to 2.2%) on each subscale, and 45 (7.7%) below the age-adjusted -2 SD level on at least one. Prevalence using these thresholds did not vary significantly with child age in either linear or non-linear (fractional polynomial) models.

*Overall agreement*

Using published norms at the -2 SD cut-point, the sensitivity of the ASQ with respect to the BSID-III was 41% and the specificity 82%. Using the distribution-based BSID-III thresholds yielded a sensitivity of 40% and a specificity of 84%. Over-referral rates (i.e., the proportion of positive screens without true delay) were 93% with published norms and 83% with the distribution-based measure. Under-referral rates (i.e., the proportion of
children negative on the screen who are delayed) were 2% with published norms and 6% for the distribution-based measure. Detailed results are shown in Table 2.

Figure 6-1. Derived -2 SD threshold for BSID-III receptive communication subscale.

Agreement for specific subscales
With published BSID-III norms, the number of children positive on each domain was very small, seriously limiting the power available to test associations. Nevertheless, agreement was significant for the gross motor (sensitivity 50%, specificity 92%) and language/communication (sensitivity 0%, specificity 96%) domains. With distribution-based thresholds, agreement was significant for cognitive/problem solving (sensitivity 25%, specificity 94%) and for gross motor functioning (sensitivity 42%, specificity 92%), but not for fine motor functioning or language/communication. Despite the fact that the
Table 6-2. Measures of agreement between ASQ and BSID-III in PANS.

<table>
<thead>
<tr>
<th></th>
<th>True -</th>
<th>False -</th>
<th>False +</th>
<th>True +</th>
<th>BSID prev.</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>PPV (95% CI)</th>
<th>NPV (95% CI)</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement with published BSID-III norms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>470</td>
<td>10</td>
<td>100</td>
<td>7</td>
<td>2.9%</td>
<td>41% (18%-67%)</td>
<td>82% (79%-85%)</td>
<td>7% (3%-13%)</td>
<td>98% (96%-99%)</td>
<td>0.02</td>
</tr>
<tr>
<td>Cognitive / problem solving</td>
<td>549</td>
<td>2</td>
<td>35</td>
<td>1</td>
<td>0.5%</td>
<td>33% (1%-91%)</td>
<td>94% (92%-96%)</td>
<td>3% (0%-15%)</td>
<td>100% (99%-100%)</td>
<td>0.17</td>
</tr>
<tr>
<td>Fine motor</td>
<td>557</td>
<td>2</td>
<td>28</td>
<td>0</td>
<td>0.3%</td>
<td>0% (0%-78%)</td>
<td>95% (93%-97%)</td>
<td>0% (0%-10%)</td>
<td>100% (99%-100%)</td>
<td>1.00</td>
</tr>
<tr>
<td>Gross motor</td>
<td>534</td>
<td>3</td>
<td>47</td>
<td>3</td>
<td>1.0%</td>
<td>50% (12%-88%)</td>
<td>92% (89%-94%)</td>
<td>6% (1%-17%)</td>
<td>99% (98%-100%)</td>
<td>0.01</td>
</tr>
<tr>
<td>Language / communication</td>
<td>557</td>
<td>4</td>
<td>26</td>
<td>0</td>
<td>0.7%</td>
<td>0% (0%-53%)</td>
<td>96% (94%-97%)</td>
<td>0% (0%-11%)</td>
<td>99% (98%-100%)</td>
<td>1.00</td>
</tr>
<tr>
<td>Agreement with distribution-based BSID-III cut-points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>453</td>
<td>27</td>
<td>89</td>
<td>18</td>
<td>7.7%</td>
<td>40% (26%-56%)</td>
<td>84% (80%-87%)</td>
<td>17% (10%-25%)</td>
<td>94% (92%-96%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cognitive / problem solving</td>
<td>542</td>
<td>9</td>
<td>33</td>
<td>3</td>
<td>2.0%</td>
<td>25% (5%-57%)</td>
<td>94% (92%-96%)</td>
<td>8% (2%-22%)</td>
<td>98% (97%-99%)</td>
<td>0.03</td>
</tr>
<tr>
<td>Fine motor</td>
<td>548</td>
<td>11</td>
<td>26</td>
<td>2</td>
<td>2.2%</td>
<td>15% (2%-45%)</td>
<td>95% (93%-97%)</td>
<td>7% (1%-24%)</td>
<td>98% (97%-99%)</td>
<td>0.12</td>
</tr>
<tr>
<td>Gross motor</td>
<td>530</td>
<td>7</td>
<td>45</td>
<td>5</td>
<td>2.0%</td>
<td>42% (15%-72%)</td>
<td>92% (90%-94%)</td>
<td>10% (3%-22%)</td>
<td>99% (97%-99%)</td>
<td>0.002</td>
</tr>
<tr>
<td>Language / communication</td>
<td>550</td>
<td>11</td>
<td>24</td>
<td>2</td>
<td>2.2%</td>
<td>15% (2%-45%)</td>
<td>96% (94%-97%)</td>
<td>8% (1%-25%)</td>
<td>98% (97%-99%)</td>
<td>0.11</td>
</tr>
</tbody>
</table>
* Fisher’s exact test.

PPV=Positive predictive value; NPV=negative predictive value; prev.=prevalence

ASQ consistently identified at least twice as many children as the BSID-III, sensitivities were uniformly poor. Detailed results are shown in Table 2. Spearman correlations of standard scores were 0.44 (p<0.01) for communication / language, 0.37 (p<0.01) for gross motor, 0.09 (p=0.02) for fine motor, and 0.16 (p<0.01) for cognitive / problem solving.

**Discussion**

In our general population sample, agreement between the ASQ and BSID-III fell short of the levels recommended for screening. In particular, sensitivity was consistently low, with more than half of cases missed. Although specificity was much better, the false positive rate remained fairly high, with five false positives for every true case. Agreement for individual domains was also poor to fair.

Although the application of published BSID-III norms to our sample led to clearly problematic prevalences, using distribution-based measures did not change the level of agreement. As we noted in a related study on the Nipissing District Development Screen (Cairney et al., 2016), this may be related to the fact that the original norms identified only the most severe cases among older children. With distribution-based thresholds, milder cases are also identified, and these are more difficult to distinguish from typically developing individuals. At the same time, the distribution-based approach should have
reduced misclassifications due to problematic norms, as well as misclassifications due to the use of age bands generally (Veldhuizen et al., 2015). These factors may have balanced the reduction in agreement resulting from the addition of milder cases.

Results of existing ASQ validation studies have varied widely. Work showing relatively good agreement with reference measures has often used high-risk or mixed samples (e.g., Hix-Small et al., 2007), which may have given rise to spectrum effects. Our results are also, however, somewhat poorer than those obtained by other large, general population studies, including that of Squires et al. (2009), Hix-Small et al. (2007), and, most notably, Limbos & Joyce (2011). The latter is (as far as we are aware) the only other large, independent, general-population validation of the English-language ASQ against standardized assessments. The reported specificity in this study was, at 78%, slightly worse than in our results, but sensitivity was considerably better, at 82%.

A comparison of our approach with that of Limbos & Joyce suggests several possible explanations for differences between our results and those of previous studies. First, Limbos & Joyce used a battery of reference instruments, while we used the BSID-III alone. Second, 28% of children were positive on the ASQ in Limbos and Joyce, implying a different effective tradeoff between sensitivity and specificity. Third, in Limbos & Joyce, parents completed the ASQ at the time of the reference assessment, and were supplied with relevant materials, where necessary. In our study, the ASQ was distributed to parents with instructions for independent completion. This probably represents a lower level of support and instruction, but may be similar to the
circumstances to be expected when the ASQ is distributed to parents in primary care and other community settings (e.g., day care, drop-in family centres). At least one study, however, has failed to find a significant effect of conditions and supporting materials on ASQ reliability (San Antonio et al., 2014).

It is also conceivable that the severe delay was comparatively rare in our sample. As Figure 1 illustrates, many individuals in our study were close to the threshold. In this setting, screening is less a question of drawing a line between two well-separated distributions than of imposing a threshold on a continuum of functioning. This can be expected to give rise to a higher error rate, because individuals can be moved across the threshold by small errors in measurement.

One specific area of concern in our data is the high false positive rate among young infants: Half of the 54 subjects receiving the 2 month ASQ (i.e., aged between 1 month and 3 months) fell below the published threshold on one or more subscales. Within this age group, caseness was also strongly correlated with exact age (R=-0.62, p<0.01), suggesting that the 2 month ASQ may cover too great a developmental period. As most prevalent developmental disorders are not identifiable among young infants in any case (First & Palfrey, 1994; American Psychiatric Association, 2013), it may be that the usefulness of standardized assessments in this age range should be reevaluated.

Significantly, young infants were excluded from the other large general-population studies we have discussed (Limbos & Joyce, 2011; Hix-Small et al., 2007; Squires et al., 1997). Removing infants under 1 year from our sample results in somewhat improved
agreement (sensitivity 50%, specificity 87%, with respect to distribution-based BSID-III thresholds).

Several other issues might also contribute to the modest level of agreement. Parent-report instruments must contend with differences in the interpretation of items, and the importance of this issue may vary from sample to sample. The use of age groups in measures like the ASQ and BSID-III can also, as the example of the 2 month ASQ illustrates, give rise to systematic misclassifications, with likely over-diagnosis among the younger children within an age band and under-diagnosis among older ones (Veldhuizen et al., 2015). There are also certain differences between the ASQ and BSID-III in terms of the domains assessed. Most notably, the ASQ includes ‘personal-social’ functioning, while the BSID-III does not. Some disagreements in status could therefore arise from differences in the implicit definitions of delay being used. Although the interval between the ASQ and BSID was generally short, this time gap may also have reduced measured agreement due to ordinary day-to-day variation in child behaviour.

Finally, reference measures are themselves not free from error. A general limitation of studies of concurrent validity is that, although they can establish whether one instrument can be substituted for another, they sometimes leave unanswered the question of whether the children identified by either are truly those who would benefit from an intervention (Streiner & Norman, 1995). Although the third edition of the BSID is perhaps too new for this question to have been examined, evidence on previous versions is somewhat mixed (e.g., Hack et al., 2005; Aylward, 2004). Recent research has often considered the
predictive validity of development screens by using formal assessments at multiple time points (Marks et al., 2008). The most important evidence on the usefulness of screening instruments, however, may come from studies examining their ability to predict practical outcomes such as special education status (e.g., Kerstjens et al., 2009) or later formal diagnosis.

Limitations
The ASQ provides a different set of items for each age group. It is possible that performance varies across these versions, but we could evaluate the measure only by aggregating them. Our sample was also fairly affluent, and socioeconomic status is associated with presence of developmental delay in children (Hackman & Farah, 2009). It is therefore possible that the true prevalence of delay was lower in our sample than in the general population. We note, however, that prevalence on the ASQ was still relatively high, and that we obtained the same results with published norms and with distribution-based thresholds. We also relied on trained research assistants, and not clinicians, to administer the BSID-III. Finally, parents completed at least two screening instruments at about the same time, and this may have increased their general awareness of developmental issues.

Conclusion
There are considerable potential benefits to the early identification of developmental delay. Our results, however, suggest that agreement between the ASQ and BSID-III is modest. Although these data must be interpreted alongside existing validation studies
and other evidence, they sound a cautionary note for general population screening with the ASQ. Effectiveness trials with clinical or functional outcomes may provide the best means of evaluating screening efforts.
References


StataCorp. 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.


CHAPTER 7: Summary, conclusions, and future research

The cause of age banding errors is simple. If a variable changes systematically with age and a single standard is applied to a range of ages, then the relative levels of functioning obtained will be wrong. This is a mathematical inevitability: Any age grouping of this kind will reduce the accuracy of the assessment. As age bands only exist when there is age-related variation, it is also reasonable to argue that systematic errors will be produced by every scoring or assessment system that uses age groups for scoring. These include grades and standardized tests in schools, performance in youth (and master’s) sports and athletics, and those assessments of functioning that use age-banded scoring. Age banding errors, then, can be expected to be very common.

Although the cause of age banding errors is simple, the errors themselves can be complex. In this program of research, I put together a simple framework for understanding these errors and then studied its implications. This led to efforts to measure age banding errors, to identify some of their less-obvious consequences, and to suggest some existing approaches that might be useful in mitigating or eliminating them.

Methods

Broadly, the analyses performed in Studies 1 through 4 include two steps. First, I use linear models to characterize age-related variation. Then, I use these results to inform a second analysis. In Studies 1 and 2, this consists simply of examining and analyzing results of the initial model; in Studies 3 and 4 of using them to guide the generation of
large, simulated datasets; and in Study 5 of using them to conduct a conventional validity analysis.

When data are cross-sectional, as they are in studies 3, 4, and 5, I use regression. In studies 1 and 2, where data are longitudinal, I use mixed effects growth curve models. In both cases, standard linear models would not provide an adequate description of age-related change – as discussed earlier, the expected mean is rarely or never a linear function of age, and the variance in functioning may itself not be constant across ages. To address this non-linearity, I use the method of fractional polynomials. This is not the most exact curve-fitting method available, but it is capable of providing a good fit to most non-linear relationships, and it produces functions that can be used in simulation or prediction efforts in a straightforward way. To take into account non-constant variance, I fit a second model examining age-related change in the residuals of the initial model. The end result of this entire modeling exercise is a pair of equations, each describing expected age-related change: One for the mean value, and one for the variance. One additional method is needed in Study 4, which considers an instrument that measures multiple developmental domains simultaneously. Here, developing an adequate description of the data requires an examination of the relationships among the different domains. To do this, I measure the ‘instantaneous’ correlations for each pair of domains, and then examine how these correlations themselves vary with age.

These results provide an abstraction of the initial data that can answer certain research questions directly, as in Studies 1 and 2. They can also, however, be used to generate
arbitrary amounts of synthetic data that exhibit the same general characteristics as the initial dataset. The resulting simulated datasets can be used to answer research questions that would otherwise be mathematically difficult or intractable, or that would require extremely large samples. These datasets support any number of numerical ‘experiments’: With an artificial sample of 1M ‘cases’, it becomes possible to impose multiple sets of age groupings and to calculate misclassification rates by simply counting records.

**Assumptions**

*Generalizability*

These results can be made arbitrarily precise, but, of course, are accurate only to the extent that the initial models are valid. The validity of these models therefore requires some examination. One consideration is generalizability. All analyses draw on three medium-sized, general-population samples from the Niagara and Hamilton regions of Ontario, Canada. Caution is therefore needed in any effort to derive expectations for any other population: Results will be directly applicable only to the extent that the population of interest resembles the one sampled. However, the included studies do not attempt to provide a means of estimating errors in any particular population, but to illustrate the errors and other consequences of age grouping. Moreover, I have avoided using external norms or other population references. Misclassification errors, for example, depend on the rate of age-related change in the mean and on the level and rate of change in the variance of the measure, not on the mean itself. Results are therefore unlikely to be very different in an otherwise similar sample that is higher- or lower-functioning.
Distributions

The remaining assumptions are distributional ones. All analyses begin from models in which ranking errors arise from the use of age bands in a normally-distributed variable. Most of these models did not assume a constant variance across age, or any particular distribution for the sample as a whole; but they did assume that the variable was normally distributed at each exact age. My analysis of banding errors in a measure with multiple domains, similarly, modeled data by assuming multivariate normality, and also assumed that inter-domain correlations would be similar at all levels of ability.

The crucial point about these distributional assumptions is that, in general, inferences do not change substantially even if the assumptions are somewhat faulty. Only a distribution lacking tails (i.e., one in which the number of observations does not decrease at more extreme values) would threaten to make the interesting results untrue, and such a distribution is implausible given the nature of things that are generally measured using measures of child development.

Although the overall pattern of results is unthreatened, the assumption of normality does affect confidence in specific numeric results. When I consider the misclassification rates of developmental assessments, for example, these rates depend on a smoothly decreasing proportion of children at decreasing levels of ability. In fact, there is some reason to imagine that this may not be quite accurate. Functioning for most individuals is likely to
result from the ordinary kind of process that produces a normal distribution – the combination of many individual variables that, in a population, can reasonably be thought of as a binomial distribution that approaches normality. The difficulty is that there are also specific syndromes that arise from particular causes, including developmental insults (e.g., fetal alcohol spectrum disorder) and genetic conditions (e.g., Angelman syndrome). The true distribution, then, is likely to be a mixture distribution – the sum of a normal distribution and a multitude of very small ones with means at low levels of functioning. The result may be a distribution with a ‘bump’ in the left tail – more people than expected at extremely low levels of functioning. As most such people would be detected by any developmental assessment, an analysis using a normal distribution might somewhat overestimate the rate of misclassifications produced by age banding.

There were three reasons for proceeding with a normal distribution despite this objection. First, although incorporating a mixture distribution into, for example, a simulation study is not technically difficult, it is very hard to know what such a distribution should look like. It should be the sum of a normal distribution and of distributions for each of many developmental syndromes with specific origins. Producing even a reasonable guess at such a distribution would clearly be very difficult, and the attempt would inevitably be a product of guesswork. Second, specific developmental syndromes are usually excluded from definitions of ‘developmental delay’ and from developmental screening efforts: Most children with them would be detected without systematic screening, and are therefore not part of the intended population for screening tests. Similarly, people with physical impairments are unlikely to be subjected to athletic evaluation in the same way
as the rest of the population. As a result, the normal distribution may, after all, be reasonably close to the population that would actually be evaluated with the types of assessments considered here. Finally, it is important to remember that this objection is strictly a theoretical one: The data underlying my conclusions do not show distributions of this type.

A related issue is relevant to the multi-domain simulation study I conducted in Project 4. The main difficulty in looking at multiple domains lay in the need to reflect associations among those domains. In this study, I took into account age-related variation in inter-domain correlations – scores on the BSID-III subtests were more weakly correlated for younger than for older children, and this had some noteworthy implications for the results and for developmental screening in general. An issue I did not address, though, is variation in inter-domain correlations with level of ability. In fact, I did not find a clear pattern of this kind in the data I had available, possibly because they included few very low or very high-functioning individuals.

**Summary of findings and commentary**

In the first project, I put forward the simple argument that relative age effects were an inevitable result of age banding. If functioning varies with age, then applying the same norm or criterion to entire age ranges will cause systematic errors. The fact that age groups exist in a given context is, in fact, a fairly certain indication that age-related errors are present: If functioning did not vary with age, age groups would be unnecessary; and if functioning varies with age and common standards are applied to age ranges, then errors
are mathematically certain. I also noted that relative age errors are not only a concern for people close to an age band limit, but will affect everyone, to a greater or lesser extent: It is only at one exact age in each group that the expected error will be zero, and, as this age can be made arbitrarily precise, the probability of anyone in a sample being that age approaches zero.

Despite the fact that age differences alone are both necessary and sufficient to explain relative age effects, many other explanations have been suggested. In Study 1, I therefore suggested a simple statistical method for deciding whether any such explanations were necessary. This involved, essentially, removing the overall age relationship from the data and then analyzing what was left. If there is still variation with relative age, then this is evidence that other explanations are needed; if there is not, then they are unnecessary. The results of my analysis implied that, as far as one simple measure of cardiorespiratory fitness was concerned, the relative age differences were fully accounted for by age itself. This does not, of course, demonstrate conclusively that no other effects operate – just that, in this dataset, their net effect is not statistically different from zero.

In interpreting the contribution of this study to research on RAEs generally, it is important to appreciate that a role for these other effects, although frequently mooted, is not always part of current models of RAEs. Instead, RAEs in athletics and academics are now often explained in terms of the “maturation/selection hypothesis”. In this view, relatively old children are more physically mature, and are therefore selected for further training and coaching. I think this view is basically accurate, and that the results of Study
I essentially support it. I did note, however, that RAEs are not a question only of biological maturation – instead, they must be the net effect of everything that varies with age, including experience and practice. In this sense, the focus on maturation per se may be inappropriate.

The second project was concerned with working out some of the implications of the simple model in Project 1. I noted, first, that increasing age is likely to shift the entire distribution of functioning upwards – i.e., the mean will increase with age, while the variance, within the short span covered by individual age bands, will stay about the same. On some hypothetical test, for example, the scores of 8 year old children might be distributed $N(50,10)$ and those of 9 year old children $N(60,10)$. An 8 year old compared to 9 year olds, then, will have his or her true level of ability underestimated by 1 SD, regardless of what this ability level is.

One of the results of this view is that, if we are interested in ranking individuals, then errors will be largest for average children and small for those far from the mean. This is because of the nature of normal (and normal-like) distributions: There are far more people close to the mean, so an absolute change of a certain size will mean “passing” a large proportion of the population. Moving below the mean in steps of 0.5 SDs, for example, means going from the 50$^{th}$ percentile to the 31$^{st}$, then the 16$^{th}$, then the 7$^{th}$, and then the 3$^{rd}$. At each step, the difference in percentile ranks is less, because there are fewer people as you move out into the tails.
If we are interested in identifying exceptional individuals, though, things are more complicated. Obviously, thresholds close to the mean will produce the largest absolute number of misclassifications. It is extreme thresholds, though, that will produce the largest age differences among the people selected. To illustrate this, we can consider two groups of children. Group 1 is exactly in the middle of an age band, while Group 2 is at the youngest limit of that band. Group 2, because they are young, have a mean that is 0.5 SD below the norm. A cut-point that is -2 SD for the age group as a whole, then, is only -1.5 SD for this group. If we apply this cut-point, we find that it identifies $\phi(-1.5) = 6.7\%$ of children in Group 2, and $\phi(-2) = 2.3\%$ in Group 1. If we use a threshold at -3 SD, though, we now get $\phi(-2.5) = 0.63\%$ in Group 2, and $\phi(-3) = 0.13\%$ in Group 1. The total proportion of Group 2 children who are wrongly identified, then, is 4.4\% in the first case, and 0.49\% in the second. Among all children identified, though, the Group 2:Group 1 ratio rises as the threshold becomes more extreme: From 2.9:1 at -2 SD to 4.6:1 at -3 SD.

This is interesting for two reasons. First, it implies that age gradients that seem small in the full population can still lead to large selection effects if the selection thresholds are extreme enough. Second, it means that we would expect to see more and more skewed age distributions at more extreme levels of functioning. We might, for example, see more relatively old people in elite sport leagues and universities than in less selective organizations. We may also see more relatively young children identified as having developmental or learning delays when more stringent thresholds are applied – a greater age difference in the “impaired” range than in the “at risk” range on a developmental assessment, for example. (As noted earlier, however, this latter example may be
complicated by the existence of specific developmental syndromes that may affect the shape of the left tail of the distribution.)

In the third project, I explored age banding issues in such an assessment. I used a dataset of real assessments to decide how the mean and variance of raw scores varied with age, and then used those results to create a large synthetic dataset. I then applied three different sets of scoring bands from real instruments and measured the misclassification rates associated with each. These proved to be meaningful (about 15%) for even the narrow bands used by a clinical assessment, and became clearly unacceptable (about 45%) with the wide bands used by one parent-completed screen.

An important secondary result from this analysis showed that the use of different age bands would prevent different measures from agreeing closely. When the bands are different, a given score will be subject to different errors on each. Someone who is old for their band on one measure and young for their band on the other, for example, would end up with quite different results on each, even if the two instruments are actually measuring exactly the same thing with no error. By comparing results for simulated individuals after applying the different sets of age bands, I was able to quantify the resulting levels of disagreement. As this analysis reflects, by design, disagreements due to age banding alone, these results represent a ceiling on the obtainable level of agreement between the measures considered.
This result is of some interest for the actual levels of agreement found, but is more important for its identification of an important problem affecting most existing developmental assessments: It is statistically impossible for them to agree closely. This may help to explain the fairly low levels of agreement reported in most validation studies, and also points to a possible source of clinical confusion – a given child can be expected to receive different scores on different measures.

In the fourth project, I extended this work to the case of developmental assessments with multiple domains. This involved modeling age-related change in means and variances for multiple subscales, and then also examining the correlation structure of the domains as a function of age. These results made it possible to then draw another large synthetic dataset. Replicating the results from Project 3 suggested that misclassification and disagreement rates with the full measure were similar to, but slightly smaller than, those from a single subscale.

These results also provided a more realistic dataset with which to explore a further issue related to the use of age bands: longitudinal instability. In the same way a child will receive different scores on measures with different age bands, he or she will also receive different scores when assessed at different points in time with the same instrument. A child who is exactly average at all ages, for example, can expect to appear above average when assessed at the upper end of an age band, but will then appear to fall sharply behind when he or she ages into the next one. Repeated assessments on an individual whose status relative to peers does not change will, in fact, show marked instability – their
passage through age bands will produce a type of saw tooth pattern, with a sudden drop as each new age band begins. This artefactual variability is clearly a problem for clinical assessment. There is also possible relevance to other areas of research, however, because repeated assessments with an age-banded measure will probably tend to exaggerate developmental discontinuity.

Finally, in the fifth project, I analyzed a validation study that compared an age banded screen (the Ages and Stages Questionnaires; ASQ) to an age banded reference measure. This study used conventional measures of agreement, including sensitivity and specificity. There were, however, two difficulties with the reference measure: It used age bands, and the published norms produced a much lower-than-expected prevalence. As concerns had been raised over these norms previously, and other measures on the same sample gave no indication that it was an unusually high-functioning one, I performed two analyses of the level of agreement. In one, I used the published norms. In the other, I derived age-specific thresholds at the notional cut-point (-2 SDs, or the -2.3rd percentile) by using quantile regression and fractional polynomials and used these to classify individuals as ‘cases’ and ‘non cases’. This analysis therefore used the sample as its own norming reference. Because the quantile regression model produced a threshold that varied continuously with age, it also removed any errors related to age banding, and therefore will have provided a fairer test of the performance of the screen. This analysis included no effort to identify or remove errors in the screen itself, as any such errors should rightly be considered part of screen as it is used; to attempt to remove them would have been to provide misleading results about the performance to be expected. (Future
analyses, however, might usefully consider attempting something along these lines in order to provide a real-world analysis of age banding errors.)

The results of both analyses were quite similar: Removing age banding errors from the reference measure did not clearly improve the performance of the screen. There are three reasons this is unsurprising. First, the analysis also involved lowering the threshold for ‘delay’. This meant that several marginal individuals were included as ‘cases’. The initial thresholds included generally more-severe cases, which should have been easier for the screen to identify. Second, the overall screen performance was quite poor: Sensitivity was approximately 40% in both analyses. At this level of agreement, most disagreements will not be due to age banding, and removing this one source of error will be very unlikely to affect the overall results. Finally, the sampling used to collect the validation sample did not produce a uniform age distribution – instead, children tended to be assessed at specific ages. This may artificially reduced the size of the age banding errors present.

This analysis does, however, demonstrate a fairly simple approach that can be used to remove age banding errors from a reference measure when the sample being assessed is believed to be similar to the population of interest. It is also a method that can be employed, when the same assumption is met, to measure age banding errors in a real dataset.
Implications of age banding

Estimating the costs of age banding errors is beyond the scope of this research, but the implications are fairly clear. First, age banding errors are likely to affect everyone who is assessed. The possibility of harm is clearly greatest for those who move across an important threshold as a result – children wrongly diagnosed with behavioral or developmental conditions, or those denied entry to advanced academic or athletic programs. In many settings, however, the smaller errors that affect large numbers of people may also be important. In academic settings, for example, it is only a few people who will gain or lose an important scholarship or team selection due to age banding errors, but essentially everyone will be affected to some lesser degree. In aggregate, these errors may be of considerable importance.

Individually, age banding errors in the measures considered here are not enormous. On any reasonable instrument, they will not be large enough to cause children with severe developmental delays to appear above average, or vice versa. They are, however, still quite important. This is particularly true in selection contexts. In part as a result of the ratio-of-tails effect discussed in Project 2, age banding is of great importance in identification of exceptional individuals – including medical or other diagnoses.

Many assessments aim to identify such exceptional people. Developmental assessments are often concerned less with locating children on a distribution than with detecting individuals with significant impairment. Schools attempt, in part, to identify students who are “gifted”, and those who would benefit from additional instruction. In sport, the
focus tends to be on identifying exceptional athletes. There are a few points that make age banding important in this ‘selection’ context.

First, the errors produced by age banding are likely to be consistent, in terms of the absolute error, across all levels of functioning. This is because increasing age will usually shift the entire distribution of functioning. This may not be wholly true – gaps for children with severe developmental syndromes, for example, may increase with age – but it is true on average for general population samples over the scales covered by individual age bands. It is implausible, for example, that the shape of a distribution would change dramatically across a 6-month or one-year age difference. This means that the error in terms of, say, z-score units will be the about the same for high-functioning, average, and low-functioning children.

Age banding errors are also important because they have the potential to create considerable confusion in clinical and research settings. Measures using different sets of age bands (and there is no general standard in this area) cannot agree closely – even if the agreement of raw scores is perfect, the use of band-specific norms means that each will assign a different level of performance to a given individual. This is likely to raise difficulties in any clinical evaluation that uses multiple measures, and it also means that any research effort to validate one measure against another will face a source of systematic error. While errors resulting from a screen’s own age bands should legitimately be counted against it, its performance will actually depend in part on how closely those bands happen to correspond with those of the reference measure. This is
likely to mean that some screens will be unfairly penalized, while the performance of others is over-estimated.

Just as importantly, age bands will create an appearance of instability over time – a child’s development will be artificially quick within an age band, and will then drop precipitously as they age into the next. A child’s apparent developmental trajectory will therefore often appear to be artificially uneven, and its exact shape will depend in large part on the exact timing of his or her assessments. This is a clear instance of a clinical illusion arising from a simple statistical problem in test development.

**Alternatives to age banding**

Sometimes, banding is close to being a logistical inevitability. There are few obvious alternatives to grouping children into grades in school, for example, and unless they are evaluated exclusively by tests that include carefully-devised scoring systems, this is likely to be reflected in the marks they receive. Sport and athletic programs, too, cannot avoid combining participants of different. In other situations, though, age bands are simply a convenience. This is generally the case with standardized instruments. There are other ways to score them, but providing norms for age groups makes it relatively easy for users.

In all cases, there is at least some scope for reducing age-related errors. Even if grouping children by age is inevitable, it may be possible to take exact age into account in some kinds of standardized assessments. Even when the amount of previous instruction or
training explains more variance than age itself, maturation may play a role; this will be
detectable statistically as a relative age effect. An adjustment on at least some
assessments that takes exact age into account has the potential to improve accuracy and
reduce misclassifications.

In clinical assessments, alternatives already exist. Although I have considered the
standard system of scoring the Bayley Scales of Infant and Child Development – and the
one invariably used in research – the assessor’s manual actually provides an alternative
means of scoring that involves plotting scores on growth charts. This may have been
intended as a means of tracking change within children rather than as a way of removing
age banding errors, but this is the effect that it has.

There are other alternative. Some short screens for developmental problems, such as the
Parental Evaluation of Developmental Status (PEDS), aim only to collect developmental
concerns. These are implicitly adjusted for exact age, as raters are expected to compare
children to others of the same age (although if the comparison is to classmates, for
example, age banding errors are not inconceivable).

Other instruments, such as the Child Development Inventory (CDI) produce a
‘developmental age’. The problem with this approach is that the difference between a
developmental and chronological age is hard to interpret, with neither the ratio nor the
difference meaningful on its own. The CDI scoring suggests that a percentage difference
between the two can be converted to a z-score, but this is very unlikely to be true,
because the pace of development changes with age. A six-month gap is insignificant for a teenager, but of some concern for an infant; while a 50% difference may be unimportant for a young infant but indicative of certain and severe delay for an adolescent.

Apart from these options, there are many statistical methods that make it possible to explore and adjust for age-related variation. These are useable with any measure that is capable of yielding an index of development – an ordinal measure of functioning that can be compared across children of different ages. Linear regression, as in Projects 1, 2, and 3, is one option, as long as a curve can be fit to the mean and distributional assumptions are met. The need for the variance to follow a particular pattern, however, makes this an approach that will be unsuited to many measures. Quantile regression, however, can usefully relate desired cut-points to age. Growth curves are also a useful approach, although they do make it difficult to score measures with a formula.

**Future research**

This work shows that age-related errors are ubiquitous and important, and that, in addition to leading to inaccurate assessments, they can lead to secondary problems such as disagreements among measures and illusions of developmental discontinuity. Several related issues, however, remain unexplored or unresolved.

First, the real-world effects of age banding in developmental assessments are still not clear. These depend on 1) how accurate instruments actually are, and 2) the populations
on which instruments are used. Age banding errors prevent measures from working as intended by systematically distorting raters’ judgements. The practical importance of these errors, however, depends on how valid the affected instruments are to begin with.

If the underlying items provide a completely accurate measure of development, then the methods used here provide a means of accurately measuring age-related errors. If the instruments are wholly inaccurate, then age-related errors are all but irrelevant (as are the measures themselves). This question has been avoided in this work partly because it cannot be answered. There are no true and accurate measures of development – indeed, there is no agreement on exactly what the developmental domains are, how they are related, or exactly how they should measured. It may, however, be possible to guess at the probable level of accuracy, or to present a range of results from which instrument users could choose depending on their own beliefs and experiences.

As noted, these results also work, in general, from normal distributions, but the existence of specific developmental syndromes may make this an inaccurate representation of the population – there may be somewhat more individuals than expected at low levels of functioning. If formal instruments are used to detect severe developmental problems, then the true errors will be somewhat smaller than these results suggest. As I have argued, however, it is likelier that severe conditions such as genetic syndromes will be detected without the use of developmental assessments; and these are, in any case, considerably less prevalent than less-severe functional problems. Nevertheless, this consideration highlights the fact that the level of error is context-dependent. Further work could therefore usefully broaden the methods used to reflect the mixture
distributions likely to be more representative of the general population, or to reflect the mix of individuals likely to be seen in, for example, pediatric clinics.

There are also some types of errors that this work has not considered. There is some interest, for example, in identifying discrepancies among developmental domains or among different kinds of functioning, in children who would not be considered to be “delayed”. The errors to be expected when comparing domains within children should be smaller than those encountered when comparing children to norms, because functioning in all areas will be over- or under-estimated for children who are, respectively, relatively old or relatively young. There are, however, still grounds to think that age banding will produce inaccuracies. First, high- or low-functioning individuals may experience increased floor or ceiling effects as they are moved closer to the lowest or highest scores attainable on an assessment, and differences among domains may thereby become harder to identify. Second, and more importantly, performance on different instrument subtests, on different academic subjects, and on different athletic tasks will change at different rates at different ages. As a result, it is quite possible for a relatively old or relatively young child to have larger errors in one area than in another. This could either mask or exaggerate real discrepancies among domains.

Finally, the question of how age-related errors can best be mitigated remains an open one. For some developmental assessments, changes in scoring offer a clear way to almost entirely resolve the problem. Tasks like the BSID could replace age-banded norms with equations, derived from growth curves or using other methods, that produce exact
estimated z-scores for every possible age. For short screens like the ASQ, changes to instrument development should make it possible to at least produce versions with a larger number of age groups, which would increase accuracy without greatly increasing the burden on parents or other users. In academic and athletic contexts, however, the grouping of children by age is probably unavoidable. Here, formal adjustments based on relative age are possible, but would be complex and controversial. Numerous other solutions have been proposed, particularly in sports and athletics, but these still require careful examination and testing. In all contexts, the optimal solution to the problem of age-related errors is still unknown.

Conclusion

In this program of research, I have tried to show how a simple logistical or mathematical convenience has led to fairly large errors in different areas of child assessment. I have explored the nature of these errors, attempted to measure them in specific contexts, and have suggested some alternative systems of scoring.

There are two reasons this work is worthwhile. One is that relative age errors lead to unfairness, misdiagnosis, and misallocation of resources. The other is that these errors are usually avoidable. All tests have error, but these are essentially a scoring artifact, which means that the accuracy of many instruments can be improved without changes to content. The result would be an improvement in validity – something that can otherwise be very difficult to achieve. The abandonment of age grouping in favour of other scoring approaches, then, could offer improved accuracy in many settings. This would result in
improved fairness in academic and athletic contexts, and in improved diagnostic accuracy in clinical ones.