Psychology
2013. Vol.4, No.3, 143-152
Published Online March 2013 in SciRes (http://www.scirp.org/journal/psych)
http://dx.doi.org/10.4236/psych.2013.43021

Schedule of Growing Skills II: Pilot Study of an Alternative Scoring Method

Margiad E. Williams[1], Judy Hutchings[1], Tracey Bywater[2], David Daley[3], Christopher J. Whitaker[4]
[1]Centre for Evidence-Based Early Intervention, Bangor University, Bangor, UK
[2]Institute for Effective Education, University of York, York, UK
[3]School of Community Health Science, Queens Medical Centre, University of Nottingham, Nottingham, UK
[4]North Wales Organization for Randomised Controlled Trials in Health & Social Care, Bangor University, Bangor, UK
Email: margiad.williams@bangor.ac.uk

Received November 26th, 2012; revised January 5th, 2013; accepted February 6th, 2013

The accurate early identification of developmental delay in young children is important. The aim of this study was to highlight and propose a solution to problems associated with scoring a UK developmental screening tool known as the Schedule of Growing Skills II. Potential problems associated with the sensitivity of this screening tool were identified. As a possible solution to this problem, an alternative scoring method was developed to yield a developmental quotient. A pilot investigation of the new scoring method was conducted through comparisons with the Griffiths Mental Development Scales. Forty-three children aged 0 - 5 years were recruited and administered both developmental assessments. Results from both assessments were compared to examine validity. Both the new and published scoring methods showed good concurrent validity, however the new scoring method demonstrated better criterion-related validity in terms of higher sensitivity, comparable specificity, generally higher over-referrals, and lower under-referrals. The Schedule of Growing Skills II could be a valid, cost-effective way of screening for developmental delay in young children using this new, more sensitive scoring method.

Keywords: Screening; Child Development; Developmental Delay; Early Intervention

Introduction

The Need for Developmental Screening

The term *developmental delay* is used to identify children that are significantly delayed in meeting developmental milestones in two or more developmental domains, with "significantly" indicating a performance of two or more standard deviations below the norm (MacDonald & Rennie, 2011). These developmental domains include motor, language, social, and academic skills. Developmental delay in children is a major problem worldwide with an estimated prevalence rate of 3% (MacDonald & Rennie, 2011). In the UK, 3% of school-aged children are identified as having a special education need associated with either a learning difficulty or an autistic spectrum disorder (National Statistics, 2012). Large numbers of children with mild or moderate learning difficulties are not detected before they enter school, despite the implementation of child health surveillance services (Mackrides & Ryherd, 2011; Hamilton, 2006). Early detection is important because studies have shown the substantial benefits that early intervention can offer children with varying disabilities (Camilli, Vargas, Ryan, & Barnett, 2010; Anderson et al., 2003).

The Use of Screening Measures

Screening tools are designed to be inexpensive, quick and easy to use to provide a snapshot that enables the identification of children needing a more thorough assessment. Some screening tools require the direct observation of a child's skills in conjunction with parental report, such as the Battelle Developmental Inventory (BDI; Newborg, Stock, Wnek et al., 1984), whilst others rely solely on parental report (Ages and Stages Questionnaire [ASQ]; Squires, Potter, & Bricker, 1999). Parental reports of child development have been shown to be one effective method of assessment for developmental delay (Glascoe, 2000; Regalado & Halfon, 2001; Sices, Stancin, Kirchner et al., 2009) and have also been shown to be considerably less expensive than developmental assessments (Hamilton, 2006).

In the UK, developmental screening is undertaken by health visitors as part of the Healthy Child Programme (HCP). The HCP provides a series of child health reviews, immunizations, screening tests, and advice and support to parents to ensure that children get the best start in life. It is the core health service for protecting, promoting, and improving the health and well-being of children (Department of Health, 2009). When a child is aged between 24 and 30 months, a health visitor may conduct a developmental check using an appropriate screening tool. The most commonly used screening tools in use in the UK are the Denver Developmental Screening Tool (DDST; Frankenberg, Fabdal, Sciarillo et al., 1981) and the Schedule of Growing Skills II (SGS II; Bellman, Lingam, & Auckett, 2008; Hall & Elliman, 2006). Such tools are considered second-level assessments within the HCP in that they are only administered to those children that have already been identified as potentially at risk using other means, e.g. by parent-report measures such as the ASQ (Squires et al., 1999) and the Parents Evaluations of

Developmental Status (PEDS; Glascoe, 1997). If children are identified by the second-level assessment as potentially at risk of developmental delay, then they are referred to a paediatrician for a more rigorous assessment using a standardised developmental assessment tool such as the Griffiths Mental Development Scales (GMDS; Griffiths, 1954, 1970).

Problems with Screening Tools

Screening is not error free but it should be as accurate as possible in order to minimise both over- and under-referrals. Some widely used screening tools, such as the DDST (Frankenburg et al., 1981) have low detection rates (Sonnander, 2000; Glascoe, 2005). Consequently, in 2006, the American Academy for Pediatrics (AAP) published recommended psychometric criteria that all developmental screening tools should meet. Specifically screening tools must have sensitivity and specificity levels of at least .70 (AAP, 2006; Hamilton, 2006). Sensitivity is the proportion of correctly identified children in need of further assessment, whilst specificity is the proportion of correctly identified children that are developing typically (Glascoe, 2005). The ASQ, a widely used screening tool, has shown sensitivity and specificity levels of .72 and .86 respectively (Squires et al., 1999), whilst the PEDS has shown sensitivity levels ranging .74 - .80 and specificity levels between .70 and .80 (Glascoe, 1997). Getting the right trade-off between sensitivity and specificity levels means that both over- and under-referral rates are minimised, which reduces the number of children incorrectly identified as either delayed (over-referral) or developing typically (under-referral). Other important characteristics for accurate screening tools are established reliability, established validity, standardisation using a large national sample, and the identification of an appropriate cut-off point (Glascoe, 2005; Sonnander, 2000).

The Schedule of Growing Skills (SGS)

The SGS II is based on Mary Sheridan's STYCAR sequences (Sheridan, 1975) and was originally developed for use in the National Childhood Encephalopathy Study (NCES) in the late 1970s. The NCES tool was designed for use with children aged between two and 36 months. Validity of the NCES tool was established by comparison with the GMDS (Griffiths, 1954, 1970), one of the few developmental assessments standardised in the UK. The NCES tool showed good concurrent validity and reliability in the form of highly significant correlations. Sensitivity levels ranged from .44 - .82 whilst specificity levels ranged from .94 - 1.0 depending on the developmental domain (Bellman, Rawson, Wadsworth et al., 1985).

Following completion of the NCES, modifications were done to make the tool simpler to use and to extend the age range to cover children from birth to five years old and its name was changed to the SGS. Since validity and reliability had already been established for the birth to three years age range, additional validity/reliability checks were only conducted for the three to five years age range. Comparisons were again carried out with the GMDS and both reliability and validity results were statistically significant (Bellman, Lingam, & Auckett, 1996) however no sensitivity/specificity calculations were conducted.

In 1996, the authors of the SGS revised the tool and conducted a standardisation in the UK. Some items were reworded, added, or removed and the developmental order of some was changed. A cognitive skills domain was also added to aid in the identification of children with cognitive deficits. Completed items related to cognitive skill, which are highlighted on the record form, are added together to give a cognitive skill score. The standardisation was conducted in England and Wales with a total of 348 children. A range of different analyses were conducted to examine item order, test reliability, and test validity. The revised SGS (SGS II) showed high levels of reliability, significant intercorrelations, and good concurrent and construct validity when compared to the DDST (Bellman et al., 1996). Age norms from the standardisation sample were also calculated and used to create the SGS II profile form which presents the age norms for each skill area. Again, no sensitivity/ specificity calculations were conducted.

The SGS II was designed to be a quick and easy tool for the developmental screening of children aged from birth to five years. It takes approximately 20 - 30 minutes for a full assessment, however since it is being used as a second-level assessment tool in the HCP, administration of individual subscales takes only a few minutes. It requires only a short course of training to use. Scoring consists of taking the score for the highest item for each subscale and transferring this score to the SGS II profile form. The child's chronological age (CA) is then added to the profile form. If the child performs within one age band of their CA, they are classed as developing typically. However, if their performance is two or more age bands below their CA, they are categorised as having possible developmental delay indicating the need for further assessment.

Problems with the SGS II

One of the main problems with the SGS II is the breadth of age bands on the profile form. The profile form consists of two-month screening developmental windows during the first year of life, but by age 18 months the developmental windows have increased to six months, and by age 36 months they have increased to 12-months wide. As a result, it may not be sensitive to developmental change as children grow older particularly since scores have to be two age bands below to be deemed as evidence of developmental delay. It also means that the developmental status of children across time, or ones of different ages, cannot be accurately compared and contrasted. This problem has been found before with a different screening tool known as the Battelle Developmental Inventory (BDI; Newborg et al., 1984). Boyd (1989) noted that normative data for the first 24 months on the BDI was presented in six-month groups and thereafter, the groups increased to 12 months. This resulted in age-related discontinuities whereby a difference of only a few days could result in a child having an average score one day and a score indicative of developmental delay the next (Boyd, 1989).

Another problem concerns the validity data for the SGS II which could potentially be flawed; performance on the SGS II was compared to performance on the DDST (Frankenberg et al., 1981), which has been shown to have low sensitivity levels (Sonnander, 2000; Glascoe, 2005), and a very small sample size was used (n = 15 for construct validity and n = 11 for concurrent validity; Bellman et al., 1996).

The Development of a Developmental Quotient (DQ)

To attempt to solve the problem associated with scoring the

SGS II, the second author devised a developmental quotient (DQ) score based on individual items that the child has completed and not the developmental windows or the highest item completed. DQs were first used by Arnold Gesell to score the Gesell Developmental Schedules (Gesell & Amatruda, 1947) and are considered an index of the current rate of development. Many subsequent developmental assessments were based on the extensive work of Arnold Gesell and have utilized DQs as a way of scoring and interpreting child developmental assessments (e.g. Griffiths, 1954; Bayley, 1969; Sheridan, 1975). When scoring the SGS II using the new scoring method, the number of successfully completed items is calculated for each skill area and this score is then converted into a developmental age (DA) score using a scoring sheet designed by the second author. The DQ for each skill area is then calculated as a ratio of the DA divided by the chronological age (CA) multiplied by 100. This method would avoid the problems associated with the current interpretation guidance based on the existing profile form.

Aims and Hypotheses

The aims of this paper were to highlight problems associated with the SGS II and to pilot an alternative scoring method. The validity of both the new and published scoring methods was compared to see whether one shows better validity than the other.

The study hypothesis is that the new SGS II scoring method will show better validity than the published scoring method in terms of consistently high sensitivity and specificity levels (ideally above .70) and low over- and under-referral rates.

Method

Participants

The inclusion criterion for this study was that the children were between birth and five years old since this is the age range of the SGS II. Children were excluded from the study if they were older than five years or if the time frame between each assessment was longer than one week. A total of 43 children were recruited from nurseries and nursery schools across North Wales. Due to limited time and resources, it was only possible to recruit a very small sample of children to undertake the study. Participants were administered the SGS II and the GMDS, respectively, at the home visit. A total of 39 (91%) of the sample completed both developmental assessments. Three were excluded on the basis of the time frame between assessments being considerably longer than one week, and one was excluded because they did not complete the GMDS assessment. The children had a mean age of 31 months (SD 11.78) with a range of nine - 52 months, and 24 (61%) of the sample were male. Two had been referred to a paediatrician due to existing developmental difficulties. All were Caucasian, 24 (62%) spoke Welsh as their first language, and 28 (72%) lived in a rural area. Some of the children showed patterns of developmental delay, according to the GMDS. Six (15%) showed delays in locomotor skills, three (8%) displayed delays in personal-social skills, five (13%) showed delays in language skills, and three (8%) displayed delays in fine motor skills (see **Table 1** for demographics).

Table 1.
Demographic characteristics of the sample.

Demographics	*n*	%
Gender		
Male	24	61.5
Female	15	38.5
Age		
0-24 months	13	33.3
25-52 months	26	66.7
Ethnicity		
Caucasian	39	100
Residence		
Urban	11	28.2
Rural	28	71.8
First Language		
Welsh	24	61.5
English	15	38.5
Present at visit		
Mother	39	100
Developmental delay[a]		
Locomotor	6	15.4
Personal-Social	3	7.7
Language	5	12.8
Fine motor	3	7.7

Note: [a]Developmental delay identified by Griffiths Mental Development Scales.

Measures

Schedule of Growing Skills II (SGS II; Bellman et al., 2008)

The SGS II is a developmental screening tool used to assess the developmental trajectories of children from birth to five years of age. It comprises ten different skill areas: passive postural (e.g. "Braces shoulders and pulls self up"), active postural (e.g. "Pulls self to stand"), locomotor (e.g. "Walks tiptoe"), manipulative (e.g. "Tower of 4 to 6 bricks"), visual (e.g. "Recognizes details of Picture Book"), hearing and language (e.g. "Follows a two-step command"), speech and language (e.g. "Names familiar objects and pictures"), interactive (e.g. "Shares toys"), self-care (e.g. "Eats skillfully with spoon"), and additional skills (e.g. "Respects the property of others"). A cognitive skills score can also be computed by adding the highlighted cognitive skill items together (e.g. "Matches all 10 colour cards") to give a cognitive skill score, however this subscale was not used in the current study. The SGS II was designed to be quick and easy to use, with administration time being approximately 20 - 30 minutes for a full assessment or shorter for a single domain assessment. A manual is provided with instructions for administering each item. Since it does not need intensive training to use, it can be used by child practitioners of varying levels of experience, including health visitors and other individuals working within a Sure Start/Flying Start Centre. A Sure

Start/Flying Start Centre provide advice and support for parents and carers and ensure that children receive the full range of services available to them from birth to five years of age.

Psychometric Properties

The SGS II was standardised in the UK in 1996 (Bellman, Lingam, & Aukett, 1996). It showed good reliability levels with an average Cronbach alpha level of .91 for internal consistency. Concurrent validity was examined using case studies of children with diagnoses and construct validity was examined by comparing the SGS II to the DDST (Frankenberg et al., 1981). However, as mentioned earlier, the validity results could be considered flawed since the DDST (Frankenberg et al., 1981) has been shown to under-detect children with developmental delay (Sonnander, 2000; Glascoe, 2005) and the sample size was very small (n = 11 for concurrent validity and n = 15 for construct validity; Bellman et al., 1996). There is no published data concerning the sensitivity and specificity of the SGS II.

Griffiths Mental Development Scales (GMDS; Griffiths, 1954, 1970)

The GMDS is a standardised tool that is used to measure the development of infants and children between birth and eight years in two versions. The birth to two years version comprises five subscales; locomotor (e.g. "Walks alone"), personal-social (e.g. "Uses spoon well"), language (e.g. "Uses 12 words"), eye-hand coordination (e.g. "Tower of 4 bricks"), and performance subscales (e.g. "Can open screw toy"). The two to eight years version has an additional practical reasoning subscale, however this subscale was not used in the current study. The scales are administered using a kit of standardised equipment and specific instructions. Administration time varies from 30 minutes to one and a half hours, depending on the age of the child being assessed. It requires a five-day extensive training to use and its use is limited to psychologists and paediatricians. The GMDS is widely used in countries including Australia, South Africa, Portugal, America and Hong Kong (Huntley, 1996).

Psychometric Properties

The GMDS is the only developmental assessment standardised in the UK. The birth to 24 months version was first standardised in the 1950s and then re-standardised in 1996 (Huntley, 1996), whilst the 24 months to eight years version was first standardised in the 1970s and then re-standardised in 2006 (Luiz et al., 2006). Average internal consistency for the birth to 24 months version subscales have been found to be .95 (Huntley, 1996) whilst Cronbach alpha coefficients for the 24 months to eight years version all exceed .70 with the average being .99 (Luiz et al., 2006). Validity information for the birth to 24 months version is not provided in the manual. For the 24 months to eight years version a facet analysis was conducted to examine the content validity of the subscales. The contents were found to be representative of their respective content domain and each item had a satisfactory degree of relevance to the construct being measured (Luiz et al., 2006).

Procedures

Nursery/Nursery School Visits

Following Bangor University School of Psychology ethical approval, a total of 12 nurseries and eight nursery schools within the counties of Anglesey and Gwynedd in North Wales were contacted to ask if they would be willing to help with parent engagement for the research. Their participation in the research involved staff giving out information packs, containing the information sheet and a cover letter explaining the study, to eligible parents. The cover letter specified whom parents should contact if they were interested in participating in the study.

Referred Children

Paediatricians in the local area were contacted via letter to ask if they were using the GMDS and if they would be willing to collaborate on the research. One paediatrician replied and was contacted to arrange a convenient time for SGS II assessments to take place. Two children had been referred to the paediatrician for GMDS assessments because of existing developmental difficulties. NHS ethics approval was given for sharing of the GMDS data and permission to administer the SGS II to both children.

Home Visit Procedure

The first author, a postgraduate student who had received training to use both the GMDS and the SGS II, conducted all the home visits. Families were visited in their home on two separate occasions. During the first home visit, parents were asked if they had read the information sheet and whether they had any questions regarding the study. If satisfied with the information and willing to participate, they were asked to read and sign a consent form. The SGS II was then administered to the child participant. Administering the GMDS first could have led to a bias in the parents' answers on the parent-report items of the SGS II because the parent had already observed the child performing these items on the GMDS assessment.

The second home visit was completed within one week of the first visit. During the second visit, the GMDS was administered to the child participant. Parents were given the option of receiving a summary report of their child's performance following the second visit. This report was based on the GMDS assessment and was checked by the second author, a Consultant Clinical Psychologist. Parents were paid £20 at the end of the second home visit for their participation.

Statistical Analyses

Each developmental tool was scored and developmental delay classified according to their manual. For the GMDS, a child was classified as having a delay if they had a DQ score below 85. For the published SGS II scoring method, the manual states that a child should be referred for further assessment if their score is two or more age bands below their CA on the profile form. For the new SGS II method, three different scores were used as the cut-off point for developmental delay, namely a DQ of less than 90, 85, or 80, to explore which cut-off point gives the most accurate results.

Statistical analyses were undertaken using SPSS version 17.0 (SPSS Inc., Chicago, IL, USA). Initial analyses determined the most accurate cut-off point for the new SGS II scoring method using Receiver Operating Characteristic (ROC) curves based on sensitivity and specificity levels. Previous studies have used this method to determine appropriate cut-off points for developmental screening tools (e.g. Meisels, Henderson, Liaw et al., 1993; Squires, Bricker, & Potter, 1997; Squires, Bricker, Heo et al., 2001). Concurrent validity was determined by correlating

developmental ages (DA) generated by each tool. Criterion-related validity was examined using 2 × 2 contingency tables to determine the concordance between classifications and by calculating sensitivity, specificity, over- and under-referral rates (see **Figure 1**). A Kappa coefficient was also calculated.

Subscale Comparisons

On the SGS II, the assessment of language is split into two skill areas, one for assessing receptive language (hearing and language) and one for assessing expressive language (speech and language). For this study, these two language skill areas were combined for comparisons with the GMDS language subscale since this subscale assesses both expressive and receptive language. Also, the assessment of personal-social skills on the SGS II is split into two skill areas, namely the interactive skill area and the self-care skill area. These were also combined in this study for comparison with the GMDS. The manipulative skills area and the visual skills area were combined on the SGS II, and for the GMDS the eye-hand coordination and performance subscales were combined. In the present analysis the composite skill is named fine motor development. The combining of subscales was undertaken to facilitate better correspondence between tools as recommended in the SGS II manual (Bellman et al., 1996). For the SGS II locomotor subscale, both the passive postural and active postural skill areas were combined with the locomotor skill area for comparison with the GMDS locomotor subscale.

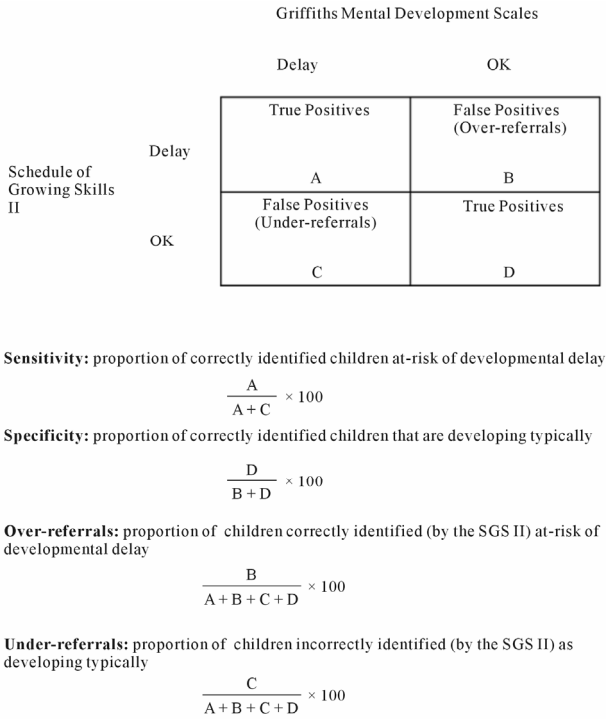

Figure 1.
Contingency table and formulas for calculating criterion-related validity. SGS II = Schedule of Growing Skills II.

Results

Age-Related Discontinuities

Table 2 shows the scores of three children within the sample who were identified by the GMDS as being significantly delayed (DQ < 70). The first child had been referred to a speech and language therapist for severe speech problems; the second and third children were the children that had been referred to a paediatrician for various developmental difficulties, including locomotor, language, and personal-social problems. According to the SGS II manual, developmental delay is represented by performing two age bands below their CA, and because the children are not yet 36 months, their scores should be placed on the 30 months age band on the profile form. The numbers in the table represent the number of age bands above/below the child's CA on the profile form.

When the children's scores are placed on the 30-month age band, the SGS II fails to identify any developmental delay for the first child, however when the scores are placed on the 36-month age band, the language delay is picked up. Similarly, the SGS II only picks up a speech delay for the second and third child on the 30-month age but picks up various developmental difficulties when placed on the 36-month age band. These results highlight the insensitivity of the profile form.

Concurrent Validity

Before commencing the data analysis, the four subscales (locomotor, language, personal-social, and fine motor) were assessed for normality for each measure. The Shapiro-Wilk test showed that three of the four SGS II and one of the four GMDS subscales were not normally distributed ($p < .05$), therefore non-parametric tests were used. Concurrent validity of both the published and new SGS II scoring methods was determined by correlations with the GMDS. DA for both SGS II scoring methods were correlated with DA generated by the GMDS using Spearman's rho. The results are displayed in **Table 3**.

Table 2.
Age-related discontinuities associated with the SGS II profile form.

CA	Age band	Loco-motor	Manipulative	Visual	Hearing & language	Speech & language	Interactive	Self-care
35	30	0	0	+1	−1	−1	0	+2
	36	−1	−1	0	−2	−2	−1	+1
33	30	−1	0	+2	−1	−3	0	−2
	36	−2	−1	+1	−2	−4	−1	−1
34	30	−1	0	+2	−1	−3	0	−1
	36	−2	−1	+1	−2	−4	−1	−2

Note: CA = chronological age; SGS II = Schedule of Growing Skills II.

Table 3.
Correlations between developmental ages for each domain.

Developmental domains	GMDS vs. SGS II (new)	GMDS vs. SGS II (published)
Locomotor	.955* $p < .001$	.934* $p < .001$
Personal-social	.950* $p < .001$	.943* $p < .001$
Language	.964* $p < .001$	.951* $p < .001$
Fine motor	.935* $p < .001$	.893* $p < .001$

Note: GMDS = Griffiths Mental Development Scales; SGS II = Schedule of Growing Skills II. *$p < .006$ (Bonferroni correction).

A Bonferroni correction was applied to the alpha level because eight correlation coefficients were tested so that the alpha level was set at $\alpha = .006$. The correlation analyses show highly significant results for both SGS II scoring methods with all correlations being significant at $p < .001$.

Establishing a Cut-Off Point

Before exploring the criterion-related validity of the two SGS II scoring methods, it was necessary to explore which DQ cut-off was the most accurate for the new scoring method. ROC curves were generated to examine the relationship between sensitivity levels and specificity levels. The true-positive rate (sensitivity) is plotted against the false-positive rate (100-specificity) for different cut-offs. The area under the curve (AUC) is a measure of test accuracy. An AUC of .5 represents an unreliable test whilst an AUC of 1 represents a perfectly reliable test. Three cut-off points were used in this analysis, namely DQ < 90, DQ < 85, and DQ < 80. The GMDS was used as the standardised assessment. ROC curves were generated by age-bands for each cut-off point across all developmental areas. **Table 4** shows the mean results from this analysis.

The ROC results show that the most accurate cut-off point for the new method of scoring the SGS II at 0 - 24 months is a DQ < 80. This cut-off gives the maximum specificity and sensitivity levels. For the 25 - 52 months age-band, the most accurate cut-off point is a DQ < 85. This is the only cut-off point showing acceptable sensitivity/specificity levels (both more than .70). For the remainder of the analyses, the new SGS II cut-off of DQ < 80 for 0 - 24 month old children and DQ < 85 for children older than 24 months will be used.

Criterion-Related Validity

Despite the correlations being highly significant for both SGS II scoring methods, some argue that correlation coefficients can be misleading (Altman & Bland, 1983; Bland & Altman, 1986). Consequently, another type of validity was calculated known as criterion-related validity. Sensitivity, specificity, over-referral rates, under-referral rates, and kappa coefficients are shown in **Table 5**.

No data was computed for the personal-social domain in the 0 - 24 months age-band because there were no children identified as having a delay in this domain. In the 0 - 24 months age-band, the new DQ scoring method shows very high levels of specificity and sensitivity with equally high kappa levels. The published scoring method shows equally high specificity levels but poor sensitivity levels with two of the four comparisons not identifying any of the delayed children. In the 25 - 52 months age-band, the new DQ scoring method again shows high specificity levels (with the exception of the locomotor domain), good sensitivity and moderate kappa levels. The published scoring method fails to identify any children with delay on three of the four comparisons but has high specificity levels.

Table 4.
Mean ROC analysis results.

Age-bands	DQ cut-off	AUC	Sensitivity	Specificity
0 - 24 months	<90	.68	.73	.63
	<85	.73	.73	.73
	<80	.77	.73	.80
25 - 52 months	<90	.80	.95	.65
	<85	.79	.78	.80
	<80	.79	.68	.91

Note: ROC = Receiver Operating Characteristic.

Discussion

The first aim of this paper was to highlight problems associated with an extensively used British screening tool known as the SGS II. Similar to the first version of the BDI (Newborg et al., 1984), this study shows age-related discontinuities associated with the SGS II profile form. The performance of children nearing their third birthday would normally be compared to the performance of 30-month old children instead of 36 months old, however this study shows that this means that developmental delay is missed and therefore those children would not be referred for further assessment. These results highlight the insensitivity of the profile form and the need to review its diagnostic usefulness. Previous studies examining this phenomenon (e.g. Boyd, 1989) suggest that instead of normative data, age-equivalent scores or similar may be more stable. This is why the second author developed a new scoring method that yields a DQ score to examine whether the SGS II could be made more sensitive to change.

The second aim of this paper was to pilot an alternative scoring method. It was hypothesised that the new DQ scoring method would demonstrate increased validity (both concurrent and criterion-related) compared to the published scoring method. Performance on the SGS II was compared to performance on a standardised developmental assessment tool (the GMDS). The overall findings show that both scoring methods show comparable concurrent validity, however the new DQ scoring method has better criterion-related validity when compared to the GMDS. The results support the study hypothesis.

The first analysis aimed to establish whether both scoring methods have good concurrent validity when compared to the GMDS. Correlation coefficients were calculated using DA and a Bonferroni correction to control for multiple comparisons. The results showed that both scoring methods showed highly significant correlations with all comparisons being significant at $p < .001$.

Table 5.
Criterion-related validity of SGS II vs. GMDS.

Age-bands	SGS II scoring	Developmental area	Sensitivity	Specificity	Over-referral %	Under-referral %	Kappa (*p*)
0 - 24 months	New DQ < 80	Locomotor	.67	1.0	0	8	.755 (.005)
		Personal-Social	-	-	-	-	-
		Language	1.0	1.0	0	0	1.00 (.000)
		Fine motor	1.0	1.0	0	0	1.00 (.000)
n = 13	Published	Locomotor	.33	1.0	0	15	.435 (.057)
		Personal-Social	-	-	-	-	-
		Language	0	.92	0	8	-
		Fine motor	0	.92	0	8	-
25 - 52 months	New DQ < 85	Locomotor	1.0	.35	58	0	.110 (.220)
		Personal-Social	.67	.88	0	4	.780 (.000)
		Language	.75	.91	8	4	.598 (.002)
		Fine motor	.50	.83	15	4	.198 (.250)
n = 26	Published	Locomotor	0	.88	0	12	-
		Personal-Social	0	.88	0	12	-
		Language	.25	1.0	0	12	.361 (.017)
		Fine motor	0	1.0	0	8	-

Note: - = data not computed due to no children being identified with developmental delay; SGS II = Schedule of Growing Skills II; GMDS = Griffiths Mental Development Scales.

The second analysis aimed to establish which cut-off score should be used for the newly developed DQ scoring method using ROC curves. This data was split into two age bands, namely children aged 0 - 24 months and children aged 25 - 52 months. The results show that the best cut-off for 0 - 24 month children is a DQ < 80 since this cut-off shows the best sensitivity/specificity trade-off. For 25 - 52 month children, a DQ < 85 was the most accurate cut-off.

The third analysis conducted explored whether the scoring methods showed good criterion-related validity. For this analysis, the data was again split into two age bands. For the 0 - 24 month sample, the new DQ scoring method showed consistently higher sensitivity, comparable specificity, lower over-referral rates, and lower under-referral rates. Kappa levels were also consistently higher and statistically significant for the new DQ scoring method. For the 25 - 52 month sample, the new DQ scoring method again showed higher sensitivity, comparable specificity, higher over-referral rates, and lower under-referral rates. Kappa levels were variable but still tended to be higher for the new DQ scoring method. The variability within the kappa levels was due to over- and under-referrals within the data. The kappa statistic does not take these levels into account and should be interpreted with caution (Altman, 1991).

In 2006, the AAP published guidelines on the recommended sensitivity and specificity levels for accurate screening. They recommend sensitivity/specificity levels of at least .70 to ensure minimum over- and under-referrals. None of the subscale comparisons for the published SGS II scoring method showed acceptable sensitivity/specificity levels across both age bands. Results for the new DQ scoring method varied across age bands. Sensitivity/specificity levels for the new DQ scoring method within the 0 - 24 month age band were high with two of the four showing acceptable levels. For the 25 - 52 month age band, only the language subscale showed acceptable levels. One reason why the levels for some subscales were not within acceptable levels could be because the sample was very small, with only 8% - 15% of the sample with a developmental delay in any one domain according to the GMDS (see **Table 1**).

The new SGS II scoring method generally had higher over-referral rates than the published scoring method giving increased sensitivity and lower specificity levels. The potential cost of high over-referral rates include the unnecessary repeated assessments with more rigorous assessment tools, and the unnecessary cost of increasing parental anxiety since parents are told their child may have a developmental delay when in fact they're developing typically (Meisels et al., 1993). However, the cost of over-referrals has been shown to be substantially less than the cost of under-referrals for both the child and society, with the cost of under-referrals being an estimated 100 times more than over-referrals (Barnett & Escobar, 1990). Additionally, Glascoe (2001) found that children with false-positive scores (or those that had been over-referred) perform significantly lower than those children with true-negative scores (those correctly identified as developing typically), and that these children might benefit from early intervention, therefore a high false-positive (or over-referral) rate is acceptable. The published scoring method had higher under-referral rates than the new method; the long-term consequences of this could be

potentially damaging to some children who would not be identified for early intervention, and may therefore develop secondary problems such as poor school performance (Anderson et al., 2003; Campbell & Ramey, 1994). There is also the issue of the cost of under-referrals to parents when a parent is told that their child is developing typically and in no need of further assessment. When their child shows increasing difficulties with everyday tasks or at school and is re-assessed, most likely following a considerable time delay, the parent is likely to feel angry or disappointed with the health system when told that their child does have a developmental delay (Meisels, 1988).

Limitations

Firstly, a very small sample was used in this study (n = 39) with the age band analyses being even smaller (n = 13 for 0 - 24 months; n = 26 for 25 - 52 months). Within the sample, only 8% - 15% had an identifiable developmental delay according to the GMDS. This could have affected the sensitivity/specificity levels and the ROC analysis results since small sample sizes may yield less precise estimates of overall diagnostic accuracy (Bachmann, Puhan, ter Riet, & Bossuyt, 2006). Also, the sample consisted of only Caucasian children who were predominantly Welsh speaking (61.5%). A larger, more diverse sample should determine whether the new DQ scoring method has consistently higher validity than the published scoring method, and whether sensitivity/specificity reach the AAP (2006) recommended levels.

Secondly, the first author who collected all of the data for the study was trained in both the SGS II and GMDS. As mentioned previously, the GMDS is only licensed for use by paediatricians and psychologists and requires a rigorous five-day training that includes sessions on child development and the development of skills to identify specific special needs that testers may come across if using the GMDS in a clinic setting (e.g. Cerebral Palsy; Autism; speech & language difficulties). The SGS II, on the other hand, only requires a one-day training, which does not include detailed information about child development. Although the SGS II is mainly used by health visitors who would have knowledge about child development, anyone working with children can complete the training. It is possible that training in the use of the GMDS may have positively influenced the way the researcher administered the SGS II due to more knowledge about child development than some users of the SGS II. The findings justify further research examining whether more knowledge regarding child development can influence the ability of the person undertaking the assessment to use and interpret screening results.

Lastly, this study does not include reliability data. The time-scale and lack of resources meant that data regarding reliability could not be collected. Future studies should examine the reliability of both the published SGS II scoring method and this newly developed DQ scoring method to determine whether one is more reliable than the other.

Implications

The implications of this study are potentially important as the SGS II is extensively used in the UK as part of the HCP. Recent Government reports recommend that all children aged 24 - 36 months should undergo a developmental check by a health visitor by increasing the coverage of the HCP to become universal (Tickell, 2011; Allen, 2011; Field, 2010). Additionally, the SGS II is being used universally as the outcome measure for an evaluation of the Flying Start Early Intervention Project across Wales which has, to date, generated data on up to 14,000 children (Welsh Government, 2009). Based on these preliminary findings, the new DQ scoring method would allow health visitors to more accurately identify those children with developmental needs than by using the published scoring method. This would lead to more children being identified swiftly and, if appropriate, getting additional required support rather than being offered support later in life when it would probably be more expensive and less effective (Allen & Duncan-Smith, 2008). The new DQ scoring method shows consistently higher sensitivity levels than the published method, which is very important considering that the SGS II is used as a second-level assessment within the HCP. The new scoring method would also make the SGS II more acceptable and useable in research practice since using a DQ score means that you can compare performances across time and across different ages.

Another implication of this study is the importance of examining different aspects of validity. Many studies exploring the validity of developmental tools have only examined concurrent validity and, therefore, used correlation coefficients as their main statistical test (e.g. Dixon, Badawi, French et al., 2009; Gollenberg, Lynch, Jackson et al., 2010; Liao, Wang, Yao et al., 2005). According to Altman and Bland (1983), using correlation coefficients can be misleading since correlation coefficients do not sufficiently highlight the variability within the data. This study is a perfect example of this. The concurrent validity data showed that both SGS II scoring methods showed highly significant correlations when compared to the GMDS. Nevertheless, when examining the criterion-related validity, the data shows that the published scoring method fails to correctly identify children with developmental delay (low sensitivity) when compared to the criterion measure (the GMDS). It is, therefore, important to explore different types of validity to ensure that the full picture is being taken into account.

Conclusion

In conclusion, this study aimed to highlight problems associated with a popular UK screening tool known as the SGS II and to pilot a new scoring method. The results show promising results in that both the published and new DQ scoring methods show good concurrent validity, however the new DQ scoring method shows better criterion-related validity in terms of consistently higher sensitivity and comparable specificity levels when compared to a standardised developmental assessment (the GMDS). Caution should be taken when interpreting these results due to the very small sample size. Based on the results of this pilot study, it is worth the cost, time, and energy to conduct a larger investigation to validate this new scoring method, which would be a useful addition to the SGS II screening tool.

REFERENCES

Allen, G. (2011). *Early intervention: The next steps. An independent report to Her Majesty's government.* London: HM Government.

Allen, G., & Duncan-Smith, I. (2009). *Early intervention: Good parents, great kids, better citizens*. London: The Smith Institute and the Centre for Social Justice.

Altman, D. G. & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician, 32,* 307-317.

doi:10.2307/2987937
Altman, D. G. (1991). *Practical statistics for medical research.* London: Chapman and Hall.
American Academy of Pediatrics [AAP], Council on Children with Disabilities, Section on Developmental Behavioural Pediatrics, Bright Futures Steering Committee, Medical Home Initiatives for Children With Special Needs Project Advisory Committee (2006). Identifying infants and young children with developmental disorders in the medical home: An algorithm for developmental surveillance and screening. *Pediatrics, 118,* 405-420. doi:10.1542/peds.2006-1231
Anderson, L. M., Shinn, C., Fullilove, M., Scrimshaw, S. C., Fielding, J. E., Normand, J. et al. (2003). The effectiveness of early childhood developmental programs: A systematic review. *American Journal of Preventive Medicine, 24,* 32-46. doi:10.1016/S0749-3797(02)00655-4
Bachmann, L. M., Puhan, M. A., ter Riet, G., & Bossuyt, P. M. (2006). Sample sizes of studies on diagnostic accuracy: Literature survey. *British Medical Journal, 332,* 1127-1129. doi:10.1136/bmj.38793.637789.2F
Barnett, W. S., & Escobar, C. M. (1990). Economic costs and benefits of early intervention. In S. J. Meisels & J. P. Shonkoff (Eds.), *Handbook of early childhood intervention.* Cambridge: Cambridge University Press.
Bayley, N. (1969). *Bayley scales of infant development.* San Antonio, TX: The Psychological Corporation.
Bellman, M. H., Lingam, S., & Aukett, A. (1996). *Schedule of growing skills II: Reference manual.* London: NFER Nelson.
Bellman, M. H., Lingam, S., & Aukett, A. (2008). *Schedule of growing skills II: User's guide* (2nd ed.). London: NFER Nelson Publishing Company Ltd.
Bellman, M. H., Rawson, N. B., Wadsworth, J., Ross, E., Cameron, S., & Miller, D. L. (1985). A developmental test based on the STYCAR sequences used in the national childhood encephalopathy study. *Child: Care, Health & Development, 11,* 309-323. doi:10.1111/j.1365-2214.1985.tb00472.x
Bland, J. M. & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 1,* 307-310. doi:10.1016/S0140-6736(86)90837-8
Boyd, R. (1989). What a difference a day makes: Age-related discontinuities and the Battelle Developmental Inventory. *Journal of Early Intervention, 13,* 114-119. doi:10.1177/105381518901300202
Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record, 112,* 579-620.
Campbell, F. A., & Ramey, C. T. (1994). Effects of early intervention on intellectual and academic achievement: A follow-up study of children from low-income families. *Child Development, 65,* 684-698. doi:10.2307/1131410
Department of Health (2009). Healthy child programme—Pregnancy and the first five years. URL (last checked October 2009). http://www.dh.gov.uk/publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_107563
Dixon, G., Badawi, N., French, D., & Kurinczuk, J. J. (2009). Can parents accurately screen children at risk of developmental delay? *Journal of Pediatrics and Child Health, 45,* 268-273. doi:10.1111/j.1440-1754.2009.01492.x
Field, F. (2010). *The foundation years: Preventing poor children be coming poor adults. The report of the independent review on Poverty and Life Chances.* London: HM Government
Frankenburg, W. K., Fandal, A. W., Sciarillo, W., & Burgess, D. (1981). The newly abbreviated and revised Denver Developmental Screening Test. *Journal of Pediatrics, 99,* 995-999. doi:10.1016/S0022-3476(81)80041-8
Gesell, A. & Amatruda, C. S. (1947). *Developmental diagnosis.* New York: Hoeber.
Glascoe, F. P. (1997). Parents' evaluations of developmental status. Nashville, TN: Ellsworth and Vandermeer Press.
Glascoe, F. P. (2000). Evidence-based approach to developmental and behavioural surveillance using parents' concerns. *Child: Care, Health and Development, 26,* 137-149. doi:10.1046/j.1365-2214.2000.00173.x
Glascoe, F. P. (2001). Are over-referrals on developmental screening tests really a problem? *Archives of Pediatric and Adolescent Medicine, 155,* 54-59.
Glascoe, F. P. (2005). Screening for developmental and behavioural problems. *Mental Retardation and Developmental Disabilities Research Reviews, 11,* 173-179. doi:10.1002/mrdd.20068
Gollenberg, A. L., Lynch, C. D., Jackson, L. W., McGuinness, B. M., & Msall, M. E. (2010). Concurrent validity of the parent-completed Ages and Stages Questionnaire, 2nd Ed. with the Bayley Scales of Infant Development II in a low-risk sample. *Child: Care, Health, and Development, 36,* 485-490. doi:10.1111/j.1365-2214.2009.01041.x
Griffiths, R. (1954). *The abilities of babies: A study in mental measurement.* London: University of London Press.
Griffiths, R. (1970). *The abilities of young children: A comprehensive system of mental measurement for the first eight years of life.* London: Child Development Research Centre.
Hall, D. & Elliman, D. (2006). Health for all children (4th ed.). Oxford: Oxford University Press.
Hamilton, S. (2006). Screening for developmental delay: Reliable, easy-to-use tools. *Journal of Family Practice, 55,* 415-422.
Huntley, M. (1996). *The griffiths mental development scales from birth to two years: Manual.* Amersham: Association for Research in Infant and Child Development (ARICD).
Liao, H., Wang, T., Yao, G., & Lee, W. (2005). Concurrent validity of the Comprehensive Developmental Inventory for Infants and Toddlers with the Bayley Scales of Infant Development II in preterm infants. *Journal of the Formosan Medical Association, 104,* 731-737.
Luiz, D. M., Barnard, A., Knoesen, N. P., Kotras, N., Horrocks, S., McAlinden, P., O'Connell, R. et al. (2006). *Administration manual of the GMDS-ER.* Amersham: Association for Research in Infant and Child Development (ARICD).
MacDonald, L. A. B., & Rennie, A. C. (2011). Investigating developmental delay/impairment. *Paediatrics and Child Health, 21,* 443-447. doi:10.1016/j.paed.2011.02.008
Mackrides, P. S., & Ryherd, S. J. (2011). Screening for developmental delay. *American Family Physician, 84,* 544-549.
Meisels, S. (1988). Developmental screening in early childhood: The interaction of research and social policy. *Annual Review of Public Health, 9,* 527-550. doi:10.1146/annurev.pu.09.050188.002523
Meisels, S. J., Henderson, L. W., Liaw, F., Browning, K., & Have, T. T. (1993). New evidence for the effectiveness of the early screening inventory. *Early Childhood Research Quarterly, 8,* 327-346. doi:10.1016/S0885-2006(05)80071-7
National Statistics (2012). Pupils with statements of Special Educational Needs (SEN) in wales, first release. SDR 88/2012. Issued by Knowledge and Analytical Services, Welsh Government. URL (last checked 26 January 2013). http://wales.gov.uk/docs/statistics/2012/120613sdr882012en.pdf
Newborg, J., Stock, J. R., Wnek, L., Guidubaldi, J., & Svinicki, J. (1984). *Battelle developmental inventory: Examiner's manual.* Allen, TX: DLMLINC Associates.
Regalado, M., & Halfon, N. (2001). Primary care services promoting optimal child development from birth to age 3 years. *Archives of Pediatric and Adolescent Medicine, 155,* 1311-1322.
Reynolds, A. J., Temple, J., Robertson, D. L., & Mann, E. A. (2001). Long-term effects of an early childhood intervention on educational achievement and juvenile arrest. *Journal of the American Medical Association, 285,* 2339-2346. doi:10.1001/jama.285.18.2339
Sheridan, M. D. (1975). *Children's developmental progress from birth to five years: The Stycar sequences* (3rd ed.). Windsor: NFER Publishing Company Ltd.
Sices, L., Stancin, T., Kirchner, L., & Bauchner, H. (2009). PEDS and ASQ developmental screening tests may not identify the same children. *Pediatrics, 124,* e640-e647. doi:10.1542/peds.2008-2628
Sonnander, K. (2000). Early identification of children with developmental disabilities. *Acta Paediatrica, 89,* 17-23. doi:10.1111/j.1651-2227.2000.tb03091.x
Squires, J., Bricker, D., & Potter, L. (1997). Revision of a parent-completed developmental screening tool: Ages and stages questionnaires. *Journal of Pediatric Psychology, 22,* 313-328.

doi:10.1093/jpepsy/22.3.313
Squires, J., Bricker, D., Heo, K., & Twombly, E. (2001). Identification of social-emotional problems in young children using a parent-completed screening measure. *Early Childhood Research Quarterly, 16,* 405-419. doi:10.1016/S0885-2006(01)00115-6
Squires, J., Potter, L., & Bricker, D. (1999). The ages and stages user's guide. Baltimore: Paul H. Brookes Publishing Co.
Tickell, C. (2011). The early years: Foundations for life, health and learning. An Independent Report on the Early Years Foundation Stage to Her Majesty's Government. URL (last checked 22 June 2011). http://www.education,gov.uk/tickellreview
Welsh Assembly Government (2009). Flying Start Guidance 2009-2010. URL (last checked 6 June 2011). http://www.wales.gov.uk/topics/childrenyoungpeople/publications/guidance0910/?lang=en