Scientific Research Publishing

Psychology, 2022, 13, 353-360
https://www.scirp.org/journal/psych
ISSN Online: 2152-7199
ISSN Print: 2152-7180

Stability Reliability of the Griffiths Scales of Child Development (3rd Edition)

Johan H. Cronje[1,2*], Elizabeth M. Green[1,2], Louise A. Stroud[1,2]

[1]Department of Psychology, Nelson Mandela University, Gqeberha, South Africa
[2]Association for Research in Infant and Child Development, Birmingham, UK
Email: *johan.cronje@mandela.ac.za

How to cite this paper: Cronje, J. H., Green, E. M., & Stroud, L. A. (2022). Stability Reliability of the Griffiths Scales of Child Development (3rd Edition). *Psychology, 13,* 353-360.
https://doi.org/10.4236/psych.2022.133022

Received: January 21, 2022
Accepted: March 14, 2022
Published: March 17, 2022

Abstract

The test-retest reliability and inter-rater reliability of the Griffiths III were investigated using a sample of 53 children in the United Kingdom and the Republic of Ireland. Correlations indicated very high reliability of the scales over a two- to four-week period (.969 to .991), as well as between raters (.967 to .996). Variations in mean scores were observed for the Gross Motor subscale, which is consistent with literature on intra-individual differences in this domain. It is recommended that the study be repeated on larger samples, at more extended testing intervals, and on samples from different countries.

Keywords

Griffiths III, Child Development, Test-Retest Reliability, Inter-Rater Reliability

1. Introduction

The reliability of test scores is an important factor in test selection (Foxcroft & Roodt, 2013). Stability reliability is an overarching concept that refers to the general stability of test scores, regardless of time or administration conditions (de Vos et al., 2014). Test-retest reliability, as an indicator of the stability of observed scores over time, is important for tests of child development as they may be used more than once on a specific child (Singh et al., 2016; Foxcroft & Roodt, 2013; Bayley, 2006). Inter-rater reliability relates to who administers or scores a test, and whether the administration and scoring instructions of a test are sufficiently robust to counteract the effect of different test administrators (Aldridge, Dovey, & Wade, 2017; Neuman, 2011). The Griffiths Scales is a test of child development and has a long history dating back to 1954 (Griffiths, 1954). A search of previous studies found the most recent mention of test-retest reliability to be

high (.82) for a previous version of the Griffiths (Huntley, 1996). No research publications about the test-retest reliability of the latest 2016 edition of the Griffiths Scales were found.

2. The Griffiths Scales of Child Development

The third edition of the Griffiths Scales of Child Development (Griffiths III) was published in 2016, following an extensive revision and builds on its predecessors as an individually administered measure of child development (Stroud et al., 2016). The first edition of the Griffiths Scales, the Baby Scales covering the first two years of a child's life, was published by Dr. Ruth Griffiths in 1954 (Stroud et al., 2016) in the United Kingdom (UK). The Griffiths Scales have been embraced by users of many countries across the globe as a developmental test for children (Sharp et al., 2018; Tso et al., 2018; Stroud et al., 2016; Cirelli, Bickle Graz, & Tolsa, 2015).

Through a reverse-reengineering process, the original theoretical understanding was found to remain a valid framework for assessing children (Stroud et al., 2016). To meet the needs of test users, however, the theory was enhanced with current assessment needs, such as executive functioning within a neurodevelopmental framework, including different forms of memory that impact general child development. The test is used in different settings which reflect its diverse user base of psychologists, paediatricians, and certain allied health professionals. The Griffiths III assesses development across the five subscales of Foundations of Learning, Language and Communication, Eye and Hand Coordination, Personal-Social-Emotional, and Gross Motor, as well as providing a General Development score. Each subscale can be administered on its own, but test administrators are advised to administer all five subscales to obtain a rounded picture of a child's abilities (Stroud et al., 2016). The age range of the test is from birth to six years.

As Griffiths III is a developmental test, items are placed in order of ascending difficulty. A child will perform items on a subscale until reaching a ceiling. At this stage, the administration of that subscale concludes, and the assessment continues with the next subscale. This format allows for multiple assessments of a child over time, as a child should face new, more difficult, test items with each administration, that are commensurate with their level of development. Test users can therefore use a single measure across multiple years of a young child's development, which creates a stable baseline whilst allowing practitioners to map a child's development. It further assists in identifying relative areas of strength and weakness for a child that allows for the creation of focussed development practices that are tailored to the specific child. The Griffiths III represents a particularly extensive update from its predecessors, so similar test-retest reliability figures cannot be assumed. The American Educational Research Association (2014) directs that each test revision seeks to improve on the validity and reliability of a test, thereby necessitating this investigation. Furthermore, given its use over time for a specific child, it is important to investigate the test's test-retest and in-

terrater reliability.

From a test-retest reliability perspective, the challenge of developmental tests is that children are expected to obtain a slightly higher raw score over time, to correspond with their chronological age and developmental trajectory (Azari et al., 2017). This upward movement in raw scores will affect some test-retest reliability scores, such as dependent-sample t-tests, thereby affecting their accuracy. T-tests are useful indicators of stability of average scores within samples, but due to the movement of such scores for developmental tests, this statistical technique tends to be absent for such tests (e.g. Azari et al., 2017; Bayley, 2006).

Additionally, if a child is assessed by different test users from time to time, it is also important to consider inter-rater reliability. In doing this confidence is created in the longitudinal profile of a child's development that may reflect the work of different test administrators over time.

Therefore, the aim of this study was to investigate the stability reliability of the Griffiths III as a function of test-retest reliability and inter-rater reliability.

3. Methods

A quantitative research method was employed to investigate the aim of the study. Approval from a National Health Service (NHS) Research Ethics Committee was not required in the UK or Republic of Ireland (ROI), as the study built on the standardisation project of the Griffiths III.

The interval between the two test sessions was two to four weeks. This interval is similar to studies for other developmental tests, such as the Bayley Scales of Infant and Toddler Development—Third Edition (Azari et al., 2017; Bayley, 2006), and the Ages and Stages Questionnaire (Singh et al., 2016).

3.1. Sample

The sample consisted of 53 children in the UK and ROI. The inclusion criterion for this study was for participants to have an uneventful medical and developmental history. The reason for this is that the study represented the first published research on the stability reliability of the Griffiths III, and therefore it sought to establish a baseline for future studies. Parents were required to complete a medical and developmental history questionnaire on their children, and this information was used to screen participants for the present study. The final sample consisted of 28 (53%) girls and 25 boys (47%). At pretest, ages ranged from three to 63 months, with a mean of 33 months (SD = 17.9). Thirty-two children were tested and retested by the same administrator, and for 21 children each test session was conducted by a different test administrator.

3.2. Data Analysis

To accommodate the expected increase in raw scores associated with child development over time, the researchers used the original standardisation sample of 426 to calculate the expected increase. This was calculated to a score of .83 to .86

per thirty days. Post-test raw scores were adjusted for each child, depending on the length of time between the pre- and post-test, and then rounded. This means that for children that were tested two weeks apart, the adjustment would not have affected post-test raw scores, whilst children who were tested three or four weeks apart had post-test raw scores reduced by one mark.

The stability of the Griffiths III scores was investigated in two ways. The first was a Pearson product-moment correlation coefficient on the pre- and post-test raw scores. To determine the significance of differences between the same and different administrators, chi-square analyses were performed, with Cramér's V for practical significance. The second set of analysis was dependent-sample t-tests to compare raw score differences.

4. Results

The descriptive statistics of scores are presented in Table 1 below. The mean scores were similar per subscale for the pretest and post-test.

For test-retest reliability, there were very high correlations for the total sample ranging from .969 to .991 between the two sets of test scores (see Table 2). The coefficients were similarly very high for the same test administrator (.967 to .989) and different test administrators (.969 to .996). Of all the scales, the difference in correlations appeared the widest for the Gross Motor subscale, but this was not statistically significant.

Dependent sample t-test was used to investigate the significance of differences in pre- and post-test raw scores for all administrators (see Table 3). The only statistically significant difference was found for the Gross Motor subscale, where a small effect of .38 was observed.

For inter-rater reliability, dependent-sample t-tests were performed for the

Table 1. Descriptive statistics for the Griffiths III pre- and post-test.

Subscale	Test	M	SD	Minimum	Maximum
Foundations of Learning	Pre	37.25	17.35	5.00	62.00
	Post	38.13	16.82	5.00	61.00
Language and Communication	Pre	40.09	18.30	4.00	63.00
	Post	40.38	17.37	5.00	63.00
Eye and Hand Coordination	Pre	39.30	17.53	5.00	65.00
	Post	40.17	16.59	6.00	64.00
Personal-Social-Emotional	Pre	41.91	17.49	6.00	64.00
	Post	42.89	17.52	7.00	64.00
Gross Motor	Pre	38.94	15.65	5.00	60.00
	Post	38.64	14.97	7.00	60.00
General Development	Pre	39.54	17.14	5.00	62.00
	Post	40.09	16.54	6.00	61.00

Table 2. Griffiths III test-retest correlation coefficients.

Subscale	All*	Same*	Different*	X^2	p (df = 1)	Cramér's V
Foundations of Learning	.979	.977	.981	0.11	.736	n/a
Language and Communication	.982	.983	.976	0.34	.560	n/a
Eye and Hand Coordination	.969	.967	.970	0.03	.854	n/a
Personal-Social-Emotional	.981	.982	.978	0.13	.719	n/a
Gross Motor	.982	.988	.967	2.71	.100	n/a
General Development	.991	.989	.996	2.69	.101	n/a

*Test Administrators.

Table 3. Differences in Griffiths III pre- and post-test scores for all administrators.

Subscale	*M*	*SD*	*t*	*p*	*d*
Foundations of Learning	.38	3.55	.77	.443	n/a
Language and Communication	−.32	3.57	−.65	.516	n/a
Eye and Hand Coordination	.32	4.39	.53	.597	n/a
Personal-Social-Emotional	.51	3.46	1.07	.288	n/a
Gross Motor	−1.13	2.99	−2.75	.008	0.38
General Development	−.09	2.42	−.26	.795	n/a

Table 4. Differences in Griffiths III pre- and post-test scores for same administrators.

Subscale	*M*	*SD*	*t*	*p*	*d*
Foundations of Learning	.91	3.92	1.31	.201	n/a
Language and Communication	−.19	3.70	−.29	.776	n/a
Eye and Hand Coordination	1.22	4.70	1.47	.153	n/a
Personal-Social-Emotional	.78	3.59	1.23	.228	n/a
Gross Motor	−.47	2.66	−1.00	.327	n/a
General Development	.53	2.73	1.09	.285	n/a

same and different administrators on pre- and post-test raw scores. No statistically significant difference was found when children had been tested and retested by the same administrator (see Table 4).

The only source of statistically significant difference was when different administrators tested children, but this difference was restricted to the Gross Motor subscale, which in turn affected General Development as an averaged score across the five individual subscales (see Table 5). These differences had a medium effect size (.66 and .69).

5. Discussion

The findings of the present study point to very high test-retest reliability of the

Table 5. Differences in Griffiths III pre- and post-test scores for different administrators.

Subscale	*M*	*SD*	*t*	*p*	*d*
Foundations of Learning	−.43	2.80	−.70	.492	n/a
Language and Communication	−.52	3.44	−.70	.494	n/a
Eye and Hand Coordination	−1.05	3.56	−1.35	.192	n/a
Personal-Social-Emotional	.10	3.28	.13	.896	n/a
Gross Motor	−2.14	3.24	−3.03	.007	0.66
General Development	−1.02	1.47	−3.17	.005	0.69

Griffiths III. The reliability figures in this study were higher than the 0.82 reported by Huntley (1996), for a previous edition of the Scales. This points to an improvement in the reliability of the Scales with the publication of the Griffiths III.

The only source of statistically significant difference appeared to be for different test administrators. The impact of reliability across different test administrators was localised to the Gross Motor Subscale, which in turn affected the General Development raw score. The Gross Motor Subscale draws on activities such as bilateral coordination, postural control, balance, power and strength, and motor sequencing (Stroud et al., 2016). Gross motor subscales are important for tests of general development to evaluate the soundness of the development within the central nervous system, and to screen for potential developmental problems (Jaščenoka, 2018; Leisman, 2016; Anderson et al., 2013). Research has also explored the relationship between motor and language development, which points to the importance of gross motor functioning in such areas as the education and learning of children (DiDonato Brumbach & Goffman, 2014; Wang et al., 2014). Referring to gross motor assessment Malina (2004) states that

> "variation in performance between testing periods probably reflects normal variation in growth (changes in body size and proportions), neuromuscular maturation, opportunity for practice, motivation to perform in the test situation, and perhaps the adults administering the tests and cooperation of young children" (p. 57).

Variability in gross motor scores is, therefore, to be expected and is more reflective of the domain than of the assessment. The observed a statistically significant difference in mean raw scores between different test administrators on gross motor performance is also consistent with the reasons provided by other authors, such as intra-individual differences (Adolph, Cole, & Vereijken, 2015; Eldred & Darrah, 2010).

6. Limitations and Recommendations

One limitation of the study is the relatively small sample size of 53 children. As the Griffiths Scales are used in several other English language speaking countries

such as Australia and New Zealand (Sharp et al., 2018), it is recommended that the test-retest reliability be investigated with samples from other countries. A further limitation is that, although the sample was appropriate for a normed developmental test and consisted of children with an uneventful medical and developmental history, the Griffiths III is most often used to confirm a deficit in the child's development, either generally or in a specific area (Cirelli, Bickle Graz, & Tolsa, 2015). So, it is recommended that the study be replicated with children that have been diagnosed with developmental delays.

The findings of this study, therefore, point to stability within the Griffiths III, as explored through test-retest and inter-rater reliability. The Griffiths III would therefore be a useful test to map the developmental trajectory of children.

Acknowledgements

Our gratitude to Dr. Danie Venter from the Nelson Mandela University for his assistance in the data analysis.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Adolph, K. E., Cole, W. G., & Vereijken, B. (2015). Intra-Individual Variability in the Development of Motor Skills in Childhood. In M. Diehl, K. Hooker, & M. Sliwinski (Eds.), *Handbook of Intra-Individual Variability across the Lifespan* (pp. 59-83). Routledge.

Aldridge, V. K., Dovey, D. M., & Wade, A. (2017). Assessing Test-Retest Reliability of Psychological Measures: Persistent Methodological Problems. *European Psychologist, 22,* 207-218. https://doi.org/10.1027/1016-9040/a000298

American Educational Research Association (2014). *Standards for Educational and Psychological Testing.* AERA.

Anderson, D. I., Campos, J. J., Witherington, D. C., Dahl, A., Rivera, M., He, M. et al. (2013). The Role of Locomotion in Psychological Development. *Frontiers in Psychology, 4,* Article No. 440. https://doi.org/10.3389/fpsyg.2013.00440

Azari, N., Soleimani, F., Vameghi, R., Sajedi, F., Shahshahani, S., Karimi, H. et al. (2017). Psychometric Study of the Bayley Scales of Infant and Toddler Development in Persian Language Children. *Iranian Journal of Child Neurology, 11,* 50-56.

Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development* (3rd ed.). Harcourt Assessment.

Cirelli, I., Bickle Graz, M., & Tolsa, J. F. (2015). Comparison of Griffiths-II and Bayley-II Tests for the Developmental Assessment of High-Risk Infants. *Infant Behavior and Development, 41,* 17-25. https://doi.org/10.1016/j.infbeh.2015.06.004

De Vos, A. S., Strydom, H., Fouche, C. B., & Delport, C. S. L. (2014). *Research at Grassroots Level* (4th ed.). Van Schaik.

DiDonato Brumbach, A. C., & Goffman, L. (2014). Interaction of Language Processing and Motor Skill in Children with Specific Language Impairment. *Journal of Speech,*

Language, and Hearing Research, 57, 158-171. https://doi.org/10.1044/1092-4388(2013/12-0215)

Eldred, K., & Darrah, J. (2010). Using Cluster Analysis to Interpret the Variability of Gross Motor Scores of Children with Typical Development. *Physical Therapy, 90,* 1510-1518. https://doi.org/10.2522/ptj.20090308

Foxcroft, C. D., & Roodt, G. (2013). *Introduction to Psychological Assessment in the South African Context* (4th ed.). Oxford University Press.

Griffiths, R. (1954). *The Abilities of Babies.* McGraw-Hill.

Huntley, M. (1996). *The Griffiths Mental Development Scales. From Birth to 2 Years.* ARICD.

Jaščenoka, J., Walter, F., Petermann, F., Korsch, F., Fiedler, S., & Daseking, M. (2018). Zum zusammenhang von motorischer und kognitiver entwicklung im vorschulalter [The Relationship between Motor and Cognitive Development in Preschool Age]. *Kindheit und Entwicklung, 27,* 142-152. https://doi.org/10.1026/0942-5403/a000254

Leisman, G., Moustafa, A. A., & Shafir, T. (2016). Thinking, Walking, Talking: Integratory Motor and Cognitive Behaviour. *Frontiers in Public Health, 4,* Article No. 94. https://doi.org/10.3389/fpubh.2016.00094

Malina, R. M. (2004). Motor Development during Infancy and Early Childhood: Overview and Suggested Directions for Research. *International Journal of Sport and Health Science, 2,* 50-66. https://doi.org/10.5432/ijshs.2.50

Neuman, W. L. (2011). *Social Research Methods. Qualitative and Quantitative Approaches* (7th ed.). Pearson International.

Sharp, M., French, N., McMichael, J., & Campbell, C. (2018). Survival and Neurodevelopmental Outcomes in Extremely Preterm Infants 22-24 Weeks of Gestation Born in Western Australia. *Journal of Paediatrics and Child Health, 54,* 188-193. https://doi.org/10.1111/jpc.13678

Singh, A., Squires, J., Yeh, C. J., Heo, K. H., & Bian, H. (2016). Validity and Reliability of the Developmental Assessment Screening Scale. *Journal of Family Medicine and Primary Care, 5,* 124-128. https://doi.org/10.4103/2249-4863.184636

Stroud, L., Foxcroft, C., Green, E., Bloomfield, S., Cronje, J., Hurter, K. et al. (2016). *Griffiths Scales of Child Development 3rd Edition; Part I: Overview, Development and Psychometric Properties.* Hogrefe.

Tso, W. Y. W., Wong, V. C. N., Xia, X., Faragher, B., Li, M., Xu, X. et al. (2018). The Griffiths Development Scales-Chinese (GDS-C): A Cross-Cultural Comparison of Developmental Trajectories between Chinese and British Children. *Child: Care, Health and Development, 44,* 378-383. https://doi.org/10.1111/cch.12548

Wang, M. V., Lekhal, R., Aaro, L. E., Holte, A., & Schjolberg, S. (2014). The Developmental Relationship between Language and Motor Performance from 3 to 5 Years of Age: A Prospective Longitudinal Population Study. *BMC Psychology, 2,* Article No. 34. https://doi.org/10.1186/s40359-014-0034-3