Title

Reliability reconsidered: Cronbach's alpha and paediatric assessment in occupational therapy

Short title

Reliability reconsidered

Category: Feature article

Journal

Australian Occupational Therapy Journal

Author

Georgia Spiliotopoulou, PhD, MSc, PGCert LTHE, BSc (Hons), FHEA

Lecturer in Occupational Therapy, School of Health Sciences and Social Care, Brunel University

Contact details

Full address:

School of Health Sciences and Social Care, Brunel University, Mary Seacole Building, Uxbridge, Middlesex, UB8 3PH, United Kingdom.

E-mail address: georgia.spiliotopoulou@brunel.ac.uk

Telephone number: (0044)18952 68827

Fax number: (0044)18952 69853

Abstract

Background / aim: Using reliable outcome measures is a necessity for the occupational therapy profession in enabling valid assessments of clients. Although Cronbach's alpha is the most widely applied index of internal consistency reliability, there are misconceptions about its use and interpretation. This paper aims to guide assessment developers in paediatric occupational therapy, as well as practitioners who are evaluating outcome measures in using and interpreting the Cronbach's alpha estimates appropriately. This will enable them to decide on the tools' clinical value and incorporate them into their practice with children.

Method: Previously published papers reporting on internal consistency issues of outcome measures in paediatric occupational therapy were searched through the Allied and Complementary Medicine database. These papers were used as a basis to discuss possible reasons for reporting of low internal consistency.

Results: The analysis demonstrates that Cronbach's alpha reports are not always interpreted in a sound way. The paper emphasises that one should be cautious about judging estimates of internal consistency. Low size of the coefficient alpha might not always indicate problems with the construction of the tool; whereas large sizes do not always suggest adequate reliability. Instead, these reports might be related to the data characteristics of the construct.

Conclusion: In judging an outcome measure's internal consistency , researchers and practitioners in occupational therapy should report and consider the nature of data, the scale's length and width, the linearity and the normality of response distribution, the central response tendency, the sample response variability and the sample size.

Key words: coefficient alpha, internal consistency, outcome measures, psychometrics, reliability

Introduction

Using outcome measures is a necessity for the occupational therapy profession, as it enables therapists to facilitate goal setting, monitor client's progress and decide on the most effective intervention (Unsworth, 2000). The use of outcome measures is also essential for evidence-based practice, which aims to the provision of the best quality health care to consumers (Unsworth, 2000). Evidence-based practice is crucial for the purpose of demonstrating the value of occupational therapy and helping the profession to gain its unique and well-deserved place among the multi-disciplinary team (Powell, 1999). A key requirement for the implementation of evidence-based practice is the ability to critically appraise outcome measures in terms of their reliability and validity, impact and applicability. This will enable occupational therapists at all levels to use the best available measures in the assessment and evaluation process and consequently provide valid and consistent information about their clients to the treatment team (Law, King & Russell, 2005).

"Toward outcome measures in occupational therapy" (Department of National Health and Welfare, and Canadian Association of Occupational Therapists, 1987) urged the incorporation of reliability reports in the developed tools. Consequently, therapists and researchers should be equipped to understand and interpret the statistical reports around reliability in determining the value, clinical utility and applicability of outcome measures (Law et al., 2005). Reliability is the ability of a tool to measure a concept in a consistent manner, and it can be assessed in various ways; therefore, researchers distinguish among "test-retest" reliability, "intrarater" and "interrater" reliability, and "internal consistency" (Hinton, 2004). This paper focuses on the "internal consistency", which refers to whether participants are responding to the different items of a questionnaire in a consistent manner in a single trial. The most sophisticated and widely applied index of internal consistency is

"Cronbach's alpha (α)". This examines the average inter-item correlation of the items in a questionnaire (Cortina, 1993). If all items are measuring the same thing (without any error) alpha will be equal to one. Otherwise, if there is no shared variance in the items, then these are supposed to reflect only "error" resulting in alpha being equal to zero (Hinton, 2004). Yet, even if alpha is close to one, this does not necessarily secure homogeneity or unidimensionality of the questionnaire (Helms, Henze, Sass & Mifsud, 2006). The reporting and correct interpretation of Cronbach's alpha is essential for judging the internal consistency of the developed outcome measures.

Judging reliability estimates

Lack of reliability is a serious drawback of an outcome measure as it indicates errors in measurements (Powell, 1999). Inconsistent outcome measures might result in invalid assessments which will consequently lead professionals to making the wrong decisions for their clients (Law et al., 2005). Conventionally, editors and reviewers consider a measure with alpha equal to or greater than 0.70 as reliable for research purposes (Bland & Altman, 1997) and this is frequently a criterion for publishing the outcome measure. But should this always be the case? Helms et al. (2006) suggested that this value is required for unspecified reasons. Furthermore, Cronbach's alpha, being a statistical tool, requires data to meet specific assumptions for the reliability estimates to be accurate and meaningful. Otherwise, the reliability of the outcome measure might be underestimated. Therefore, Pedhazur and Schmelkin (1991) proposed that the reported reliability should be evaluated by taking into account the specific circumstances of each study before claiming lack of reliability for a developed outcome measure.

One has to be cautious about judging reliability estimates. Ottenbacher (1995) and Ottenbacher and Tomchek (1993) have already discussed that authors in therapeutic research use inadequately the statistical tools and misinterpret the statistical results of interrater, intrarater, and test-retest reliability. No published studies have been identified on the use and interpretation of internal consistency reports in therapeutic research, although literature presents concerns about its use in the field of psychology. Moreover, there are concerns that occupational therapists may sometimes evaluate the research findings in a problematic way (Nutley & Davies, 2000). This along with the lack of published resources in occupational therapy to provide strategies for reporting and analysing internal consistency estimates to promote sound interpretation, make the publication of such guidelines imperative.

This paper aims to guide potential assessment developers in paediatric occupational therapy, as well as practitioners and reviewers in interpreting the alpha estimates reported in outcome measures. This will enable them to evaluate these measures and decide whether they are suitable to be used in practice. For this purpose, previously published papers reporting on the internal consistency of outcome measures in paediatric occupational therapy are discussed. Following that, there are guidelines on determining appropriate use of Cronbach's alpha and ways to evaluate whether the internal consistency of a tool may be underestimated or overestimated.

Cronbach's alpha in paediatric occupational therapy assessment

Studies reporting on the internal consistency of outcome measures in paediatric occupational therapy were searched through the Allied and Complementary Medicine database. The search used a combination of the key words "occupational therapy", "children", and "internal consistency". Papers published between 2000 and 2008 were selected. The

search identified 8 papers satisfying the above criteria. All of them used Cronbach's alpha to assess the internal consistency of the outcome measures. Five of these papers (see table 1) indicated that there were issues with the internal consistency of the tools they were researching. Therefore, these papers were selected to discuss possible reasons related to the reporting of problematic internal consistency. These issues are discussed followed by suggestions on sound reliability reporting and evaluation.

The number of items included in an outcome measure is implicated in the interpretation of internal consistency estimates. Katz, Golstand, Bar-IIan and Parush (2007) reported on the internal consistency of "The Dynamic Occupational Therapy Cognitive Assessment for Children", which consists of 56 items divided in 5 cognitive subtests. The reported Cronbach's alpha estimates for these subtests ranged between 0.61 and 0.77. Although, the researchers suggested that the internal consistency of the outcome measure was moderate to high, 2 of the 5 subtests fell below the benchmark of 0.70 which usually determines acceptable reliability. The below 0.70 subtests were (a) "orientation" consisting of 8 items ("α" = 0.61), and (b) "visuomotor construction" consisting of 7 items ("α" = 0.61). The researchers suggested that probably the small number of items in each subtest resulted in these "relative moderate coefficients". Indeed, it has been shown that Cronbach's alpha estimation of reliability increases with scale length (i.e. number of items in the scale) (Cronbach, 1951; Voss, Stem & Fotopoulos, 2000). Yet, Swailes and McIntyre-Bhatty (2002) suggested that the effect on alpha is particularly noticeable when the number of items is below seven. In the outcome measure of Katz et al. (2007), the number of items included in both subtests is equal to or above the critical number of seven. Therefore, small number of items cannot explain the moderate coefficients in this tool indicating that there might be issues with the construct or with certain items.

Further to the above Cronbach (1951) provided the following correction factor to account for the acknowledged sensitivity of alpha to scale length. This formula is an estimate of the mean inter-item correlation (ρ) and is independent of scale length:

$$\rho = \frac{\alpha}{n-(n-1)\alpha} \qquad (1)$$

where:

ρ = an estimator of reliability independent of scale length,

α = coefficient alpha, and

n = the number of items in the scale.

Although the above formula has been little used (Voss et al., 2000), it would be very useful for researchers, reviewers and practitioners in evaluating the internal consistency of an outcome measure. By calculating the mean inter-item correlation (ρ), which is independent of the scale length, one would be able to evaluate the internal consistency of a tool by comparing the size of this mean inter-item correlation (ρ). Values of the mean inter-item correlation (ρ) vary widely with the topic area under investigation and the nature of research, but seldom exceed 0.50 (McKennell, 1978). Clark and Watson (1995) recommended a mean inter-item correlation (ρ) within the range of 0.15 to 0.20 for outcome measures that measure broad characteristics (i.e. general constructs such as extraversion) and between 0.40 and 0.50 for those tapping narrower ones (i.e. specific constructs such as talkativeness). Based on formula (1) the mean inter-item correlation (ρ) of the "orientation" subtest was 0.16, and of the "visuomotor construction" 0.18.

Formula (1) would also be useful for evaluating the internal consistency of outcome measures with a satisfactory alpha estimate and a large number of items. As Cronbach's alpha

increases with the number of items in a test, one could raise the reliability estimates by increasing the number of items in the test. However, this does not reflect best practice as tests tend to be extremely long without always guaranteeing internal consistency or unidimensionality (McKennell, 1978). For example the reported alpha estimate for the "praxis" subtest in the test of Katz et al. (2007) was 0.70. However, this subtest consisted of 23 items, a fact that would inevitably result in a high alpha estimate. Based on formula (1), the mean inter-item correlation (ρ) of this subtest was 0.09, which is lower than the mean inter-item correlation (ρ) of the "orientation" and "visuomotor construction". This indicates lower internal consistency of "praxis" in comparison to the "orientation" and "visuomotor construction" subtests.

The width of a scale is another factor which influences the interpretation of reliability estimates. Katz et al. (2007) did not clarify what was the nature of the data derived from each subtest of their instrument, or the way that each item was scored. By looking at the aggregated possible scoring for each subtest and the number of items this includes, it seems that for "spatial perception", "orientation", "praxis", "visuomotor construction", and "thinking operations" the possible score range was 1 to 2, 0 to 2, 0 to 2, 1 to 5, and 1 to 5, respectively. Thus, the width of the scale for "orientation" (for which alpha was below 0.7) was quite limited (3-points' scale). Voss et al. (2000) suggested that wider scales may have a greater variance, which should increase alpha and that this variation has been found to happen in scales with over 4-points' width. Therefore, the small width scale (3-points) might be one possible explanation for the low alpha estimate of the "orientation" subtest.

Further to the width of scales, Voss et al. (2000) found that scales with a central point (e.g. 5-points) tend to have a higher alpha estimate in comparison to scales with an even

number of points (e.g. 6-points). This is defined as “central response tendency”; still, it is unclear whether offering the respondents the opportunity to take a middle position encourages more honest, consistent and reliable responses, or whether allows them to avoid making decisions and stating their opinion.

The nature of data is also important in interpreting reliability reports. In the outcome measure of Katz et al. (2007), although that was not explicitly stated, the data of “spatial perception” seemed to be of nominal nature (possible score range: 1 to 2). Kuder- Richardson (K-R 20) is a statistical tool which is considered to be more appropriate for estimating the internal consistency of outcome measures with nominal data in comparison to the use of Cronbach’s alpha (Carey, 1994). However, Katz et al. (2007) did not report using the Kuder-Richardson (K-R 20) formula. Similarly, Brown and Gaboury (2006) calculated Cronbach’s alpha for the “Test of Visual-Perceptual skills – Revised” using 356 children aged 5 to 11 years. For each age group and for each subtest separately, the 37 out of the 49 reported alpha estimates were below the benchmark of 0.70. Brown and Gaboury (2006) indicated that clinicians should consider that this outcome measure does not measure reliably children’s visual-perceptual skills across different age levels. However, this test comprises of 7 visual-perceptual subtests, each one including 16 items, which are scored as 0 to 1 (nominal data). Therefore, the reliability reports for both of the above outcome measures might have been underestimated due to the use of Cronbach’s alpha; whereas the Kuder-Richardson (K-R 20) would be a more appropriate statistical tool (Carey, 1994).

The sample size might also influence reliability estimates. Klein, Sollereder and Gierl (2002) identified reliability issues when they calculated Cronbach’s alpha with 294 children aged 6 to 12 years for the unrevised version of the “Test of Visual-Perceptual skills”. The

alpha of each subtest for each age group ranged between 0.23 and 0.89. Yet, the authors attributed the low alpha levels to the small sample size of each group. Lane and Ziviani (2003) have also explained the low alpha ("α" = 0.40) in one of the 10 subtests of the "Test of Mouse Proficiency" as a result of small sample size. Indeed, maximizing the number of participants responding to a scale can increase the value of alpha by increasing the amount of covariance among item responses (Helms et al., 2006). Still, Helms et al. (2006) indicated that small samples can also provide large reliability coefficients and that there is a debate around what is an "appropriate sample size" for calculation of reliability. Therefore, they suggested the conduction of reliability power analyses for anticipated sample sizes.

The variability of data is another factor that should be considered for interpreting internal consistency reports. May-Benson and Koomar (2007) assessed the internal consistency of the "Gravitational Insecurity" outcome measure. The alpha of the total test score was 0.717 for 18 children with gravitational insecurity aged 5 to 10 years and 0.479 for their matched typically developing children. The authors attributed the low estimate for the latter group to low variability in the data. Helms et al. (2006) suggested that reliability is driven by variance with greater scores variance leading to greater score reliability. Hence, one would expect that a more heterogeneous sample (as a group of typically developing children) should yield higher reliability estimates in comparison to a more homogeneous group (as a group with children with gravitational insecurity) on a measure of gravitational insecurity. Low alpha estimates for the diagnosed sample would reflect that the scale is functioning as it should (Helms et al., 2006). Still, for the above test the typically developing children yield the lower alpha. Therefore, if this low alpha cannot be explained by other factors, then it might indicate a problem with the construction of the measure.

Normal distribution and linearity of data are important prerequisites for the use of Cronbach's alpha. Yet, none of the above studies discussed characteristics of data such as linearity and normality. The formula from which Cronbach's alpha derives means that the coefficient alpha is equal to the reliability of an outcome measure only when the subtests' or items' true scores are linearly related (Zimmerman, Zumbo & Lalonde, 1993). Moreover, Wilcox (1992) showed that Cronbach's alpha is sensitive to even minor deviations from normality, which is due to a heavy tail's effect that greatly influences the estimation of the variance. Yet, heavy tails are a common occurrence in psychometric measurement. Zimmerman et al. (2003) suggested that researchers pay relatively little attention to the consequences of violating these two assumptions. They also proposed that the discrepancies between coefficient alpha in the sample and the population reliability coefficient, which represents the coefficient alpha in the entire target population, are likely to be large when these assumptions are not met.

Implications for assessment in occupational therapy and guidelines for interpreting internal consistency estimates

Considering the above, researchers and practitioners should be cautious when evaluating internal consistency estimates to decide upon the value of an outcome measure used in occupational therapy. It is not always theoretically sound to divide outcome measures as reliable or unreliable based on rigid benchmarks (i.e. the 0.70 benchmark) (Voss et al., 2000). In some occasions, the reliability of measures used in occupational therapy may be underestimated by the current formulas used for calculation of Cronbach's alpha when the data do not meet the assumptions of normality and linearity, or when the data are of nominal nature (Pedhazur & Schmelkin, 1991; Voss et al., 2000). In other cases, the reliability reports may be underestimated due to the limited number of items included in the test, or due to the

limited width of the scale used to measure these items (McKennell, 1978; Voss et al., 2000). Yet, in some other occasions, the reliability of the test might be overestimated because of the inclusion of a large number of items.

Therefore, the present paper suggests that researchers, reviewers and practitioners should consider the following guidelines for interpreting internal consistency estimates:

1. Check that the statistical tool is appropriate for the level of measurement of data. For nominal data, the Kuder-Richardson (K-R 20) formula should be used instead of Cronbach's alpha.
2. Check that the data are normally distributed and linear. If not, then Cronbach's alpha will underestimate the reliability of the outcome measure.
3. Check the scale's length and decide whether the reported alpha is adequate. To make such a decision, calculate the mean inter-item correlation (ρ) which is independent of scale length by using formula (1). Although values of the mean inter-item correlation (ρ) vary widely with the topic area and the nature of research, they seldom exceed 0.5. A recommended mean inter-item correlation (ρ) for instruments that measure broad characteristics falls within the range of 0.15 to 0.20 and between 0.40 and 0.50 for those tapping narrower ones.
4. Check the width of the scale. Scales of less than 4-points width might result in underestimated alpha reports.
5. Check whether there is central response tendency. Scales of over 4-points width with a central point (e.g. 5-points) may have a higher alpha estimate in comparison to scales with an even number of points (e.g. 6-points). Aiming for central response tendency is not clear yet as to whether it reflects good practice.

6. Check the sample size. Larger samples may increase the alpha estimates. Researchers are advised to conduct reliability power analysis for anticipated sample sizes, as there is no definite rule as to what is an appropriate sample size.
7. Consider variability of data. More heterogeneous samples (as a group of typically developing children) should yield higher reliability estimates in comparison to a more homogeneous group (as a group of children with a specific impairment) on a measure of this specific impairment. Lower alpha estimates for the diagnosed sample would reflect that the scale is functioning as it should.
8. Researchers should provide clear information on the data characteristics derived from the outcome measures under investigation.

Conclusion

The use of Cronbach's alpha in occupational therapy research should not be done in a perfunctory way, but rather should reflect informed decision making about which set of measurement assumptions one's data best fit. Also, for some outcome measures because of the data characteristics of the construct and with our present state of knowledge, researchers, practitioners and reviewers should think that they should probably accept lower figures of alpha estimates rather than the conventionally set benchmark of 0.70. In cases where an alternative formula can be applied (e.g. Kuder-Richardson formula for nominal data), then researchers should ensure that they do so. Researchers, practitioners and reviewers should also consider whether it is appropriate to accept outcome measures with a high alpha estimate when the number of items included in the scale is too large.

Calculating and reporting reliability coefficients for outcome measures are appropriate and good practices. Nevertheless, it is the author's responsibility to also provide the necessary

information regarding data's characteristics to enable the reader to critically evaluate the results and judge the value of the tool. It is also the practitioner's and reviewer's responsibility to comprehend the reported values and interpret them in a broader frame of rigorous research before welcoming the outcome measure in the clinical world.

Acknowledgments

The author would like to thank Dr Anita Atwal and Ms Nicola Plastow, lecturers in occupational therapy, Brunel University for their valuable comments on the manuscript.

References

Bland, J.M. & Altman, D.G. (1997). Statistics notes: Cronbach's alpha. *British Medical Journal*, 314, 572.

Brown, G.T. & Gaboury, I. (2006). The measurement properties and factor structure of the test of visual-perceptual skills-revised: Implications for occupational therapy assessment and practice. *American Journal of Occupational Therapy*, 60(2), 182-193.

Carey, L.M. (1994). *Measuring and evaluating school learning*, 2nd ed. Boston: Allyn and Bacon.

Clark, L.A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7, 309-319.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 96-104.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

Department of National Health and Welfare, and Canadian Association of Occupational Therapists (1997). *Toward outcome measures in occupational therapy*. Ottawa, ON: Department of National Health and Welfare.

Helms, J.E., Henze, K.T., Sass, T.L. & Mifsud, V.A. (2006). Treating Cronbach's alpha reliability coefficients as data in counseling research. *The Counseling Psychologist*, 34(5), 630-660.

Hinton, P.R. (2004). *Statistics explained*, 2nd ed. London: Routledge.

Katz, N., Golstand, S., Bar-IIan, R.T. & Parush, S. (2007). The dynamic occupational therapy cognitive assessment for children: A new instrument for assessing learning potential. *American Journal of Occupational Therapy*, 61(1), 41-52.

Klein, S., Sollereder, P. & Gierl, M. (2002). Examining the factor structure and psychometric properties of the test of visual-perceptual skills. *Occupational Therapy Journal of Research*, 22(1), 16-24.

Lane, A. & Ziviani, J. (2003). Assessing children's competence in computer interactions: Preliminary reliability and validity of the test of mouse proficiency. *OTJR: Occupation, Participation and Health*, 23(1), 18-26.

Law, M., King, G. & Russell, D. (2005). Guiding therapists decisions about measuring outcomes in occupational therapy. In: M. Law, C.M. Baum & W. Dunn (Eds), *Measuring occupational performance: Supporting best practice in occupational therapy*, (2nd ed. pp. 33-44). Thorofare, NJ: SLACK Inc.

May-Benson, T.A. & Koomar, J.A. (2007). Identifying gravitational insecurity in children: A pilot study. *American Journal of Occupational Therapy*, 61(2), 142-147.

McKennell, A. (1978). Attitude measurement: Use of coefficient alpha with cluster and factor analysis. In: J. Bynner & K.M Stribley (Eds), *Social research: Principles and procedures*. Milton Keynes: Open University Press.

Nutley, S. & Davies, H.T.O. (2000). Making a reality of evidence-based practice: Some lessons from the diffusion of innovations. *Public Money and Management*, 20, 35-42.

Ottenbacher, K.J. (1995). An examination of reliability in developmental research. *Journal of Developmental and Behavioral Pediatrics*, 16(3), 177-182.

Ottenbacher, K.J. & Tomchek, S.D. (1993). Reliability – Analysis in therapeutic research – Practice and procedures. *American Journal of Occupational Therapy*, 47(1), 10-16.

Pedhazur, E. J. & Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Powel, N.J. (1999). Research principle used in developing assessments in occupational therapy. In: B.J. Hemphill-Pearson (Ed), *Assessments in occupational therapy mental health: An integrative approach*, (pp.341-350). USA: SLACK Inc.

Swailes, S. & McIntyre-Bhatty, T. (2002). The "Belbin" team role inventory: Reinterpreting reliability estimates. *Journal of Managerial Psychology*, 17(6), 529-536.

Unsworth, C. (2000). Measuring the outcome of occupational therapy: Tools and resources. *Australian Journal of Occupational Therapy*, 47(4), 147-158.

Voss, K.E., Stem, D.E., Jr. & Fotopoulos, S. (2000). A comment on the relationship between coefficient alpha and scale characteristics. *Marketing Letters*, 11(2), 177-191.

Wilcox, R.R. (1992). Robust generalizations of classical test reliability and Cronbach's alpha. *British Journal of Mathematical and Statistical Psychology*, 45 (November), 239-54.

Zimmerman, D.W., Zumbo, B.D. & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Education and Psychological Measurement*, 53, 33-49.

Tables

Table 1

Published papers reporting on issues related to the internal consistency of outcome measures used in paediatric occupational therapy.

Name of outcome measure	**Authors**	**Journal**	**Year**
The test of visual-perceptual skills	Klein, Sollereder and Gierl	Occupational Therapy Journal of Research	2002
The test of mouse proficiency	Lane and Ziviani	OTJR: Occupation, Participation and Health	2003
The test of visual-perceptual skills-Revised	Brown and Gaboury	American Journal of Occupational Therapy	2006
The dynamic occupational therapy cognitive assessment for children	Katz, Golstand, Bar-IIan and Parush	American Journal of Occupational Therapy	2007
Gravitational insecurity assessment	May-Benson and Koomar	American Journal of Occupational Therapy	2007