*European Journal of Psychological Assessment*, doi: 10.1027/1015-5759/a000206

**ARTICLE**

# Equivalence of paper and computer formats of a child self-report mental health measure

Praveetha Patalay,[1,2] Jessica Deighton,[1] Peter Fonagy,[2] and Miranda Wolpert[1]

[1] Evidence Based Practice Unit, University College, London, and Anna Freud Centre, London, United Kingdom,

[2] Department of Clinical, Educational and Health Psychology, University College London, United Kingdom

Correspondence to: Praveetha Patalay
Evidence Based Practice Unit (EBPU),
University College London and the Anna Freud Centre,
12 Maresfield Gardens, London NW3 5SU
E-mail: praveetha.patalay.11@ucl.ac.uk

## Abstract

Research examining the equivalence of paper and computer-based adult mental health measures has found mixed results, and this issue has not been explored for child self-report measures. Results from adult studies cannot be generalised to young people, especially taking into consideration research indicating that current generations are more comfortable disclosing sensitive information on computer-based media. This paper investigates the psychometric equivalence of the paper (*N* = 777) and computer (*N* = 777) formats of a child and adolescent self-report mental health measure, 'Me and My School' (M&MS), completed by school pupils aged 8–14 years. Common practice in equivalence testing has been to use scale-level analysis and factor structure equivalence; the limitation being inability to assess format-based differences at the item-level. We conduct differential item functioning (DIF) analysis to assess whether item-response probability is different based on survey format. Results demonstrate that young people completing the M&MS on paper have lower scale-level overall scores. However, DIF analyses indicate that this difference is not explained by item-level probabilities. The results suggest that survey format equivalence testing of other widely used child and adolescent mental health measures may be necessary before data from different formats are directly compared or combined.

***Key Words:*** computer survey, psychometric equivalence, mental health, children, DIF

## Introduction

Computers are increasingly being used with adult and child populations to complete questionnaires, whether for population-based epidemiological surveys, assessing health outcomes in services or screening for problems. Computer-based survey methods are recognised as having many benefits over paper survey methods, such as increased efficiency of data collection and management and reduced coding errors, which in turn increase the speed at which feedback and results can be produced (Hayslett & Wildemuth, 2004; Kays, Gathercoal, & Buhrow, 2012). Even though computer based surveys have many advantages, paper based surveys may sometimes be preferable, especially in clinical settings and in settings where access to computers is limited.

However, studies of questionnaire format have found that format influences survey response rates (Hayslett & Wildemuth, 2004), item response and missed items, especially for items of a sensitive nature (Kays et al., 2012), and social desirability effects (Booth-Kewley, Larson, & Miyoshi, 2007). This indicates that psychometric equivalence between different survey formats, such as paper-based and computer-based, cannot be assumed.  In support of this, the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999) highlighted the need for cross-format equivalences to be established prior to direct comparison of data collected from paper-based surveys and computer- and internet-based surveys. The Scientific Advisory Committee of the Medical Outcomes Trust (2002) included validation of alternate forms of measures into their assessment framework for measures of health outcomes and quality of life. Hence, it is critical to establish when and where there is cross-format equivalence before data collected from multiple formats can be considered comparable.

While psychometric equivalence is not routinely tested in mental health measurement, especially in child and adolescent mental health, some widely used adult mental health measures have been investigated, yielding mixed results. Holländare, Andersson, and Engström (2010) tested the equivalence of the Beck Depression Inventory (BDI-II) and the Montgomery–Asberg Depression Rating Scale – self-rated and found partial format effects for the BDI-II.. Wijndaele et al. (2007) assessed the equivalence of five mental-health-related measures in adults, including the General Health Questionnaire (GHQ-12) and Symptom Checklist (SCL-90-R). Equivalence varied depending on the measure, with low test–retest coefficients for the SCL-90-R and high coefficients for the GHQ-12. Whitehead (2011) tested the equivalence of internet and paper formats of the Hospital Anxiety and Depression Scale and the Fatigue Symptom Inventory using separate samples, and found significantly higher fatigue being reported online.

While different formats of many of the most widely used child self-report mental health measures exist, such as the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) and the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Rescorla, 2001), equivalence testing is not common practice. In seeking to establish equivalence of these different formats for children, it is not sufficient to generalise from findings from adult measures. There are many differences in how young people experience computer-based environments compared with adults (Prensky, 2001). Current generations of young people are not only more comfortable and proficient with computers and newer technologies, but they also use them to self-express and are more comfortable disclosing sensitive information on these media (Livingstone, 2008; Turner et al., 1998). The proposed explanation is that young people’s conception of privacy and intimacy when using computers and the internet are different from that of previous generations, mainly as a result of having grown up with access to these technologies (Livingstone, 2008).

Existing research testing equivalence between survey formats consistently uses scale-level approaches such as means, internal consistency, correlations, and testing constancy of factor structure across formats. These statistics are necessary but not sufficient to establish equivalence of survey formats as they provide information only at the scale-level. Scale-level analyses do not account for how individuals at different levels of the latent construct perform on the individual items of the instrument (Hambleton & Jones, 1993; Raykov & Marcolides, 2011). This limitation is dealt with in item-level approaches, such as IRT, as at the core of these approaches is a model that describes how individual subject responses on items of an instrument relate to an unobservable trait (Hambleton & Jones, 1993; Raykov & Marcolides, 2011).

One of the key constructs within the item-based framework is looking at item response probability based on different groups, which is termed differential item functioning (DIF;

Walker, 2011). DIF occurs when individuals at the same level of the trait or construct being measured have unequal probabilities of attaining a given score on a given item, usually on the basis of sociodemographic grouping such as gender or ethnicity (Rogers, 2005). DIF analysis therefore attempts to disentangle these item-performance differences while controlling for overall score on the latent trait. We propose using item-level analysis such as DIF alongside scale-level analyses to establish whether differences exist at the item-level, the scale-level or both.

In mental health measurement, especially child and adolescent mental health measurement, techniques such as DIF analysis are not yet widely used (Sharp Goodyer & Croudace, 2006), even though they have been in existence for a few decades and are routinely used in measure construction and evaluation in other fields, particularly education (e.g. Le, 2009). More recently, measures of child and adolescent mental health have used item-level analyses  either to justify existing scale properties and items (Sharp et al., 2006) or to help in item selection (Deighton et al., 2013; Ravens-Sieberer et al., 2007). Petersen et al. (2003) used DIF analysis to compare translations of an emotional functional scale (EORTC QLQ-C30). Extending this application of DIF, the methodology can be applied to assess whether there are differences in item functionality across different formats (e.g., paper and computer) of the same instrument.

In light of the mixed results found with adult measures of mental health, and literature highlighting the differences between current generations of young people and adults in terms of the relationship they have with computers and the internet, we propose to test the psychometric equivalence of the computer and paper survey formats of a self-report measure of mental health, the Me and My School Questionnaire (M&MS; Deighton et al 2013), that was developed as a community based screening-tool of general mental health for children as young as 8 years old. The M&MS has also recently been included as a routine outcome measure in specialist mental health services in the UK. As far as the authors are aware, this is the first study looking at equivalence between the paper and computer-based formats of a child and adolescent self-report measure of general mental health.

In terms of study methodology used to test psychometric equivalence between survey formats, two approaches are commonly used: same sample repeated measurement (e.g. Holländare et al., 2010) and comparison of separate demographically similar samples (e.g., Ritter et al., 2004). In this study we have chosen to use demographically matched samples to avoid practice and order effects associated with repeated measurement (as found in Holländere et al., 2010).

This paper seeks to establish whether there is psychometric equivalence between the paper and computer-based versions of a child self-report mental health measure, the M&MS , both at the scale-level and at the item-level, with the expectation that, given that children might be more comfortable disclosing information on a computer-based medium, there might be psychometric inequivalences.

## Method

### Sample

Participants were school-aged children from school years 4 to 9 (age 8–14 years) in the English school system. Sociodemographic information collected included gender, ethnicity and eligibility for free school meals, which is often used as a proxy for socioeconomic deprivation in school-based research in England (Hobbs & Vignoles, 2010).

Paper surveys. Paper survey data were collected from three secondary schools and four primary schools across England. A total of 863 pupils from seven schools answered the paper version of the M&MS questionnaire. Out of these, 777 (90%) participants completed all the items from both the emotional and behavioural difficulties scales. For the item-level DIF analysis all items needed to be complete, and for this reason only pupils who completed all items were included in the analysis.

Computer surveys. The computer-based survey comparison group (N = 777) was selected from a large comparison pool of 39,168 pupils (comprising 87.3% of the total sample with no missing items) from 630 primary and 180 secondary schools who completed the survey as part of a national study of mental health in schools (Wolpert et al., 2011). Exact matching on all four demographic characteristics (year group, gender, ethnicity and socioeconomic status) was possible and was carried out using psmatch2 (Leuven & Sianesi, 2003) in STATA (StataCorp, 2011) to create a one-to-one matched comparison group. This method is akin to a random allocation approach and ensures that any differences between the two groups are not due to differences in such key demographic predictors (Bland & Altman, 1994).  After taking into account the demographic variables the subsample of 777 computer-based surveys was not different from, and hence was representative of, the large pool of

39,168 pupils who completed the computer-based survey (emotional difficulties, β = 0.002, p = .72; behavioural difficulties, β = 0.005, p = .35).

As the samples were exactly matched, there were absolutely no differences in the sample characteristics of both groups on measured socio-demographic variables. The pupils in both samples ranged from Year 4 to Year 9 (7.3% Y4, 15.4% Y5, 7.1% Y6, 11.8% Y7, 33.7% Y8, 24.6% Y9). 51.7% of the sample was male (n = 402). 53.7% of the sample was classified as White British followed by 16.9% Black, 10.4% White other, 9% Mixed, 7.1% Asian, 1.9% other, and 1% of participating pupils were unclassified. 23% (n = 179) of the sample were eligible for free school meals (FSM). In terms of representativeness, the sample was more deprived than the national school population (FSM: sample 23%, national 12–14%) and had a lower proportion of White British pupils (sample 53.7%, national 73–77%; Department for Education, 2010).

## Me and My School questionnaire (M&MS)

The M&MS (Deighton et al., 2013) is a 16-item measure comprising a 10-item emotional difficulties scale and a 6-item behavioural difficulties scale. Participants respond to each item by selecting one of three options: Never, Sometimes, Always. The items in the measure do not exhibit DIF on the basis of demographic groupings such as having English as an additional language, special educational needs and socioeconomic status (Deighton et al., 2013). The measure has at-risk thresholds with a score of 10 and above (10-11 borderline, 12 + clinical) indicating problems on the emotional difficulties scale and 6 and above (6 borderline, 7+clinical) indicating behavioural problems on the behavioural difficulties scale (Deighton et al., 2013). Given the focus on item-level analysis in the current paper Table 1 presents item response descriptives for the paper and computer survey samples.

<insert Table 1 around here>

## Procedure

Paper surveys were completed in classroom-based sessions during the normal school day in the presence of class teachers and facilitated by researchers to ensure privacy of responses. Computer-based surveys were completed by pupils using computers in school within the normal school day with support from their class teachers. Pupils could access their questionnaire with a unique code that was assigned to them by the research team. Both online and paper versions presented items in a clear and child-friendly manner and in exactly the same sequence; the key difference between the formats is that in the computer version items were presented one at a time, whereas in the paper version they were presented one below the other. In both phases of data collection consent was sought from parents via mail beforehand. All pupils received information about the study, including explanation of the confidentiality of their responses and their right to decline to participate and drop out at any time.

## Analysis and Results

Five steps were taken to establish the level of equivalence between the paper and computer-based versions of the measure. First, scale-level mean comparisons were carried out for the emotional and behavioural difficulties scales reported via paper and computer-based formats. Second, a categorical data confirmatory factor analysis was conducted to confirm whether the known existing factor structure of the measure fitted the data collected from paper questionnaires. Third, internal reliabilities of the scales in both formats were compared using the Cronbach's alpha coefficient of internal consistency. Fourth, DIF analysis was conducted to assess the equivalence of item-response probabilities in the two formats. Lastly, Differential Test Functioning (DTF) analysis was conducted to compare how the entire set of items functioned in the different survey formats

### Means and variances

Overall, as can be seen in Table 2, mean scores were significantly lower in the paper survey sample compared with the computer survey sample for both emotional difficulties (effect size, d = 0.2) and behavioural difficulties (d = 0.24). Owing to the large age range in the current sample, further analyses were done separately in the primary school aged children (8-11 years) and the secondary school aged adolescents (11-14 years) to assess if age-specific differences were present. This division by age is also of interest because, as mentioned in the introduction, most self-report measures of mental health have been designed and validated for children aged 11 years and above. As can be seen from Table 2, the format based differences were found in both the younger and older participants.

<insert Table 2 around here>

Confirmatory factor analysis (CFA)

A categorical data CFA was conducted in Mplus 6 (Muthén & Muthén, 2007) to ascertain whether the known factor structure (Deighton et al 2013) fitted the data from completed paper surveys (Table 3). The model was specified such that cross-loading between items and factors did not occur and the two scales were treated as two unidimensional scales. The fit indices (CFI = 0.96; TLI = 0.95; RMSEA = 0.06) indicated good model fit based on widely accepted criteria of model fit (Hu & Bentler, 1999). CFA co-efficients for the computer based survey are also presented in Table 3 to allow for comparisons (CFI = 0.95; TLI = 0.95; RMSEA = 0.06).

Internal reliability

Cronbach's alpha in the paper surveys (emotional difficulties, α = 0.78, behavioural difficulties, α = .81) and the computer surveys (emotional difficulties, α = 0.80, behavioural difficulties, α = .82) were adequate and similar for corresponding scales in both formats.

DIF

DIF analysis to determine if any of the items operated differently based on survey format can be done using a variety of approaches including IRT, Mantel- Haenszel, logistic regression and the Rasch model (Karami, 2012). Two common approaches were used here, 1) Liu–Agresti common log odds ratio (L-A-LOR; Liu & Agresti, 1996) which is based on the -Haenszel common-odds ratio generalised to polytomous data and represents the log odds ratio of one group selecting a response option compared with the other group when the level of the overall measured construct is the same, and 2) IRT approach with graded response model (Samejima,1997) was used and the presence of DIF is indicated by the difference in model fit estimates on the X^2 distribution between the model with both the item difficulty and discrimination parameters constrained to be equal and the model where they are allowed to be estimated freely. L-A-LOR was estimated in DIFAS 5.0 (Penfield, 2005) and IRT based DIF model was estimated in IRTPRO (Paek & Han, 2012). The size of the DIF was interpreted using a widely accepted classifying system whereby DIF in polytomous items is considered negligible if L-A-LOR < 0.43, moderate if between 0.43 and 0.64, and large if >0.64 (Penfield, 2007).

< insert Table 3 around here>

Table 3 presents the L-A-LOR and X^2difference values for all the items in the two scales. One item in the emotional difficulties scale, ‘Nobody likes me’, exhibited a statistically significant but negligible DIF based on the criteria outlined above (Penfield, 2007). Two items in the behavioural difficulties scale, ‘I am calm’ and ‘I break things on purpose’, also exhibited statistically significant yet negligible amounts of DIF. The negative L-A-LOR of ‘I am calm’ indicates that it was easier to endorse in the paper format, whereas the positive L-A-LOR value for ‘I break things on purpose’ indicates that it was easier to endorse in the computer based format of the measure. The results obtained via the IRT approach largely cross-validated the results found based on the L-A-LOR approach. The only discrepancy was that the IRT approach did not find significant DIF for ‘I am calm’ in the behavioural scale.

Additionally, DIF analysis by age-group (primary and secondary school aged participants), indicated similar patterns in items that exhibited DIF in the two age-groups on the behavioural difficulties scale. However, for the items in the emotional difficulties scale only younger participants exhibited significant and moderate DIF on the item ‘Nobody likes me’ (L-A-LOR= -.56), whereas older participants did not exhibit significant DIF on any of the items.

DTF

DTF assesses the aggregate effect of DIF across all the items in a scale (Penfiled & Algina, 2006) and was analysed using the v2 statistic in DIFAS 5.0 (Penfield, 2005). Co-efficients are presented in Table 2. Based on criteria for assessing the size of DTF (Penfield & Algina, 2006) a v2 <.07 is considered negligible and hence the DTFs were deemed not to warrant concern.

## Discussion

Formal equivalence testing of different formats of measures used to assess child and adolescent mental health is not yet a widely adopted practice, and even though many of the most widely used measures are available in both paper and computer-based formats, not much is known about their equivalence across formats. The current paper aimed to, first, test the psychometric equivalence of a child self-report mental health measure, ‘Me and My School’ and, secondly, demonstrate how DIF analysis can be used to assess item-level differences alongside current methodologies used to assess scale-level differences.

Our results indicate that there are overall differences in mean scores for the paper and computer-based version of the M&MS questionnaire and the DIF analysis indicates that format differences at the item-level are almost non-existent, except for one item which displayed moderate DIF only in younger children. However, the DTF analysis in both the overall and age-specific samples suggested the effect of DIF across all the items was negligible. The discrepancy between the scale-level psychometric inequivalences and the item-level equivalences suggests that the difference in scores between the formats is due to an overall 'dampening' of scores in the paper format. This might be attributed to differences in the level of disclosure to topics of a sensitive nature, such as mental health, in the different formats, or to differences in the characteristics of the two methods in terms of perceived privacy and confidentiality afforded by the survey formats. These points are related, as they both reflect the level of participants' comfort with and likelihood to disclose information based on the survey medium. They indicate that the increase in use of technology and social networking sites by young people might influence their readiness to disclose sensitive information via computer and internet-based media (Livingstone, 2008; Turner et al., 1998).  Our results so far indicate that young people might be more comfortable disclosing information on sensitive issues such as their mental health on computer-based measures than paper-based ones. This difference should be explored further in terms of young people's levels of comfort with the different formats and the effects of different levels of format familiarity on item response.

The DIF analyses carried out in the current paper illustrate how format-based differences can be assessed at item-level. The fact that the scale-level analysis and the item-level analysis lead to different conclusions about the equivalence of the M&MS measures emphasises the importance of using both methods. We suggest that future studies looking to establish equivalence between formats should also use this item-level analysis alongside complete scale or measure-level analyses to gain greater understanding of where there are psychometric equivalences and inequivalences between different formats of the same measure. Additional properties of measures, such as sensitivity to change, might also benefit from similar explorations in future research.

While the current study marks a step forward in methodological approaches to evaluating equivalence in child mental health measures, the main limitation is that pupils completing the questionnaires were not randomly allocated to the paper or computer survey conditions and the allocation to different formats was at the school level. While this is a consideration, in England, less than 3% of variation in mental health scores is explainable at school level (Wolpert et al, 2011), which suggests that the results are not attributable to allocation at the school level. Although the proportions of pupils missing items in both the paper and the computer versions were similar, the current study does not explore in-depth the possibility of differences in missing items and their differential predictors in the two samples.

The results of this study raise the question of how to deal with psychometric inequivalence across survey formats when it exists. In the case of the M&MS, as differences were mainly at the scale-level it would be possible to account for format inequivalence in other analysis. Further research with this measure is required to assess whether these results are replicated and whether the amount by which the paper-based surveys result in lower scores remains consistent across different samples and settings. As computer-based survey administration is likely to become more common, this is an area of study that could benefit from more research and discussion.

Additionally, this study uses a recently developed and validated measure which has not been subjected to as much psychometric scrutiny. Although the current study adds to the psychometric understanding of this measure, the generalisability of the results to other widely used measures is limited. However the current results indicate that it is necessary to test equivalence of formats for other measures of mental health before using different formats widely and interchangeably. The study raises concerns about how measures are currently used across all settings without having been tested sufficiently for equivalence. Even though the effect sizes of the difference for the M&MS measure would be considered small (Cohen, 1988), where non-random allocation of formats occurs this could have a significant impact on the outcomes of intervention studies. Moreover, when used for screening the differences in proportions identified as being at-risk is large enough to warrant concern when used at a population level. Until further research is done to assess the psychometric equivalence of commonly used child mental health measures across formats, some degree of caution is warranted when combining or directly comparing data collected via different formats and in repeated measurement studies.

## References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles.* Burlington, VT: University of Vermont Research Center for Children, Youth and Families.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Bland, J. M., & Altman, D. G. (1994). Matching. *British Medical Journal, 309*, 1128. doi: /10.1136/bmj.309.6962.1128

Booth-Kewley, S., Larson, G. E., & Miyoshi, D. K. (2007). Social desirability effects on computerized and paper-and-pencil questionnaires. *Computers in Human Behavior, 23*(1), 463-477. doi: 10.1016/j.chb.2004.10.020

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Deighton, J., Tymms, P., Vostanis, P., Belsky, J., Fonagy, P., Brown, A., . . . Wolpert, M. (2013). The Development of a School-Based Measure of Child Mental Health. *Journal of Psychoeducational Assessment*, 31, 247-257. doi:10.1177/0734282912465570

Department for Education. (2010). *Schools, Pupils and their Characterictic*.  London: Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/s000925

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38*, 581-586. doi: 0021-9630/97

Hambleton, R. K., & Jones, R. W. (1993). Comparison of Classical Test Theory and Item Response Theory and their Applications to Test Development *ITEMS*: NCME.

Hayslett, M. M., & Wildemuth, B. M. (2004). Pixels or pencils? The relative effectiveness of Web-based versus paper surveys. *Information Science Research, 26*(1), 73-93. doi: 10.1016/j.lisr.2003.11.005

Hobbs, G., & Vignoles, A. (2010). Is children's free school meal 'eligibility' a good proxy for family income? *British Educational Research Journal 36*(4), 673-690. doi: 10.1080/01411920903083111

Holländare, F., Andersson, G., & Engström, I. (2010). A Comparison of Psychometric Properties Between Internet and Paper Versions of Two Depression Instruments (BDI-II and MADRS-S) Administered to Clinic Patients. *Journal of Medical Internet Research, 12*(5). doi: 10.2196/jmir.1392

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Karami, H. (2012). An Introduction to Differential Item Functioning. *International Journal of Educational and Psychological Assessment, 11*, 59-76.

Kays, K., Gathercoal, K., & Buhrow, W. (2012). Does survey format influence self-disclosure on sensitive question items? *Computers in Human Behavior, 28*(1), 251-256. doi: 10.1016/j.chb.2011.09.007

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. International Journal of Testing, 9(2), 122-133. doi:10.1080/15305050902880769

Leuven, E., & Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Boston College Department of Economics.

Liu, I. M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*(4), 1223-1234.

Livingstone, S. (2008). Taking risky opportunities in youthful content creation: teenagers use of social networking sites for intimacy, privacy and self-expression. *New Media & Society, 10*(3), 393-411. doi: /10.1177/1461444808089415

Muthén, L. K., & Muthén, B. O. (2007). Mplus User's Guide (6th edition). Los Angeles, CA.

Paek, I., & Han, K. T. (2012). IRTPRO 2.1 for Windows. *Applied Psychological Measurement*. doi: 10.1177/0146621612468223

Penfield, R. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement,, 29*(2), 150-151. doi: 10.1177/0146621603260686

Penfield, R. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20*(3), 335-355. doi: 10.1080/08957340701431435

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*(4), 295-312.

Petersen, M. A., Groenvold, M., Bjorner, J. B., Aaronson, N., Conroy, T., Cull, A., . . . Sullivan, M. (2003). Use of Differential Item Functioning Analysis to Assess the Equivalence of Translations of a Questionnaire. *Quality of Life Research, 12*(4), 373-385.

Prensky, M. (2001). Digital Natives, Digital Immigrants: Part 1. *On the Horizon, 9*(5), 1-6.

Ravens-Sieberer, U., Gosch, A., Rajmil, L., Erhart, M., Bruil, J., Power, M., . . . the European KIDSCREEN Group. (2007). The KIDSCREEN-52 Quality of Life measure for children and adolescents: Psychometric results from a cross-cultural survey in 13 European Countries. *Value in Health, 11*(4), 645–658. doi: /10.1111/j.1524-4733.2007.00291.x

Raykov, T., & Marcolides, G. A. (2011). *Introduction to Psychometric Theory*. New York: Routledge.

Rogers, H. J. (2005). Differential Item Functioning. *Encyclopedia of Statistics in Behavioral Science*: John Wiley & Sons, Ltd.

Samejima, F. (1997). Graded response model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* N.Y.: Springer.

Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing Health Status and Quality of Life instruments: Attributes and Review Criteria. *Quality of Life Research, 11*, 193-205. doi: http://www.jstor.org/stable/4038039

Sharp, C., Goodyer, I. M., & Croudace, T. (2006). The Short Mood and Feelings Questionnaire (SMFQ): a unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. *Journal of Abnormal Child Psychology, 34*(3), 379-391. doi: 10.1007/s10802-006-9027-x

StataCorp. (2011). Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent Sexual Behavior, Drug Use, and Violence: Increased Reporting with Computer Survey Technology. *Science, 280*(5365), 867-873. doi: 10.1126/science.280.5365.867

Walker, C. M. (2011). What's the DIF? Why Differential Item Functioning Analyses Are an Important Part of Instrument Development and Validation. *Journal of Psychoeducational Assessment, 29*(4), 364-376. doi: 10.1177/0734282911406666

Whitehead, L. (2011). Methodological Issues in Internet-Mediated Research: A Randomized Comparison of Internet Versus Mailed Questionnaires. *Journal of Medical Internet Research, 13*(4), e109. doi: 10.2196/jmir.1593

Wijndaele, K., Matton, L., Duvigneaud, N., Lefevre, J., Duquet, W., Thomis, M., . . . Philippaerts, R. (2007). Reliability, equivalence and respondent preference of computerized versus paper-and-pencil mental health questionnaires. *Computers in human behavior, 23*(4), 1958-1970. doi: 10.1016/j.chb.2006.02.005

Wolpert, M., Deighton, J., Patalay, P., Martin, A., Fitzgerald-Yau, N., Demir, E., . . . Frederikson, N. (2011). Me and my school: findings from the national evaluation of Targeted Mental Health in Schools. Nottingham: DFE. Retrieved from: https://www.education.gov.uk/publications/RSG/publicationDetail/Page1/DFE-RR177

Table 1

*Item response proportions in the paper and computer survey formats*

| Item | Item response % | | | | | |
|---|---|---|---|---|---|---|
| | Paper survey | | | Computer survey | | |
| | Never | Sometimes | Always | Never | Sometimes | Always |
| *Emotional difficulties scale* | | | | | | |
| I feel lonely | 64.7 | 33.7 | 1.5 | 61.1 | 34.7 | 4.1 |
| I cry a lot | 69.0 | 29.5 | 1.5 | 65.1 | 30.6 | 4.2 |
| I am unhappy | 40.9 | 57.5 | 1.5 | 38.0 | 58.9 | 3.1 |
| Nobody likes me | 64.1 | 33.1 | 2.8 | 66.7 | 28.3 | 5.0 |
| I worry a lot | 39.9 | 53.7 | 6.4 | 37.3 | 52.8 | 9.9 |
| I have problems sleeping | 65.6 | 28.7 | 5.7 | 58.7 | 31.5 | 9.8 |
| I wake up in the night | 41.6 | 50.5 | 8.0 | 36.9 | 49.3 | 13.8 |
| I am shy | 41.6 | 51.4 | 7.1 | 38.6 | 52.6 | 8.8 |
| I feel scared | 65.4 | 33.5 | 1.2 | 57.4 | 39.4 | 3.2 |
| I worry when I am at school | 68.2 | 28.4 | 3.3 | 61.1 | 33.5 | 5.4 |
| *Behavioural difficulties scale* | | | | | | |
| I get very angry | 42.3 | 50.2 | 7.5 | 35.9 | 51.7 | 12.4 |
| I lose my temper | 46.6 | 45.8 | 7.6 | 40.2 | 48.2 | 11.7 |
| I hit out when I am angry | 64.7 | 27.7 | 7.6 | 55.1 | 33.2 | 11.7 |
| I do things to hurt people | 79.2 | 19.6 | 1.3 | 73.0 | 23.2 | 3.9 |
| I am calm | 5.4 | 60.2 | 34.4 | 7.3 | 60.6 | 32 |
| I break things on purpose | 87.9 | 11.2 | 0.9 | 80.7 | 14.9 | 4.4 |

Table 2

*Scale-level means and DTF statistics for the overall sample and sub-samples broken down by age*

| | | ***Overall sample*** | | ***8-11 years (primary school sample)*** | | ***11-14 years (secondary school sample)*** | |
|---|---|---|---|---|---|---|---|
| Scale | Statistic | Paper survey | Computer survey | Paper survey | Computer survey | Paper survey | Computer survey |
| *Emotional Difficulties Scale* | M (SD) | 4.78(3.23) | 5.4(3.58) | 5.21(3.37) | 6.17(3.31) | 4.60(3.15) | 5.16(3.64) |
| | t-test (df) | t(1552) = 3.93*** | | t(462)=3.08** | | t(1065.88)=2.74** | |
| | % above threshold | 9.3% | 13% | 12.1% | 16.4% | 8.1% | 11.6% |
| | DTF $v^2$ (SE) | .01(.01) | | .02(.03) | | -.003(.01) | |
| *Behavioural Difficulties Scale* | M (SD) | 2.75(2.35) | 3.35(2.65) | 2.23(2.25) | 3.24(2.44) | 2.97(2.36) | 3.39(2.73) |
| | t-test (df) | t(1552) = 4.72*** | | t(462)=4.64*** | | t(1064.9)=2.68** | |
| | % above threshold | 11.7% | 19.4% | 9.5% | 19% | 12.7% | 19.6% |
| | DTF $v^2$ (SE) | .03(.03) | | .01(.05) | | .03(.03) | |

***p<.001, **p<.01

Table 3

*CFA standardized loadings and DIF co-efficients*

| | **CFA** | | | | **DIF** | |
|---|---|---|---|---|---|---|
| | Paper survey | | Computer survey | | L-A-LOR (SE) *(focal group=Paper)* | $X^2$difference (df=2) *(IRT approach)* |
| Item | I | II | I | II | | |
| *Emotional difficulties scale* | | | | | | |
| I feel lonely | .67 | | .63 | | -.01(.13) | 4.9 |
| I cry a lot | .69 | | .68 | | .01(.13) | 4.4 |
| I am unhappy | .66 | | .73 | | -.1(.13) | 0.8 |
| Nobody likes me | .56 | | .64 | | -.37*(.13) | 12.9* |
| I worry a lot | .75 | | .68 | | -.14(.12) | 2.2 |
| I have problems sleeping | .58 | | .62 | | -.17(.12) | 3.4 |
| I wake up in the night | .50 | | .59 | | -.14(.11) | 5.4 |
| I am shy | .34 | | .29 | | .00(.12) | 0.3 |
| I feel scared | .83 | | .72 | | .17(.14) | 1.8 |
| I worry when I am at school | .78 | | .76 | | .09(.13) | 0.3 |
| *Behavioural difficulties scale* | | | | | | |
| I get very angry | | .93 | | .87 | -.05(.14) | 1.3 |
| I lose my temper | | .94 | | .83 | -.07(.15) | 0.7 |
| I hit out when I am angry | | .81 | | .85 | .23(.15) | 0.7 |
| I do things to hurt people | | .71 | | .79 | .07(.26) | 2.3 |
| I am calm | | .62 | | .58 | -.37*(.13) | 1.9 |
| I break things on purpose | | .60 | | .66 | .38*(.17) | 10.3* |

*Note.* 1.All loadings in CFA are significant at $p$ < .001. * statistically significant at the 0.05 level. 2. Negative L-A-LOR values indicate DIF favouring the focal group i.e. for the same level of construct easier to endorse for the focal group. Conversely, positive L-A-LOR values indicate the item is more difficult to endorse for the focal group.