UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Larsen, S. A., Asbury, K., Coventry, W. L., Hart, S. A., Little, C. W., & Petrill, S. A. (2023). Measuring CHAOS? Evaluating the Short-form Confusion, Hubbub and Order Scale. *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.77837

Methodology and Research Practice

Measuring CHAOS? Evaluating the Short-form Confusion, Hubbub and Order Scale

Sallly A. Larsen[1][a], Kathryn Asbury[2], William L. Coventry[3], Sara A. Hart[4,5], Callie W. Little[5], Stephen A. Petrill[6]

[1] School of Education, University of New England, Armidale, Australia, [2] Department of Education, University of York, York, UK, [3] School of Psychology, University of New England, Armidale, Australia, [4] Department of Psychology, Florida State University, Tallahassee, FL, USA, [5] Florida Center for Reading Research, Florida State University, Tallahassee, FL, USA, [6] Department of Psychology, The Ohio State University, Columbus, OH, USA

Keywords: home environment, twin studies, child development, measurement invariance, scale validation, confirmatory factor analysis

https://doi.org/10.1525/collabra.77837

Collabra: Psychology

Vol. 9, Issue 1, 2023

The Confusion, Hubbub and Order Scale (CHAOS) – short form – is a survey tool intended to capture information about home environments. It is widely used in studies of child and adolescent development and psychopathology, particularly twin studies. The original long form of the scale comprised 15 items and was validated in a sample of infants in the 1980s. The short form of the scale was developed in the late 1990s and contains six items, including four from the original scale, and two new items. This short form has not been validated and is the focus of this study. We use five samples (*N*=10,898) from studies in Australia, the UK, and the USA, to examine the measurement properties of the CHAOS short form. We first compare alternate confirmatory factor models for each group; we next test between-group configural, metric and scalar invariance; finally, we examine predictive validity of the scale in each sample under different conditions. We find evidence that a two-factor configuration of the six items is more appropriate than the commonly used one-factor model. Second, we find measurement non-invariance across groups at the metric invariance step, with items performing differently depending on the sample. By contrast we find longitudinal measurement invariance in two of the three samples with multi-wave data collection on the CHAOS. Finally, we report inconsistent results in tests of predictive validity using family-level socioeconomic status and academic achievement as criterion variables. The results caution the continued use of the short-form CHAOS in its current form and recommend future revisions and development of the scale for use in developmental research.

The effect of home environments on childhood functioning and development has been a topic of research interest for decades (Bradley, 2015; Evans, 2006). Bronfenbrenner's widely-known bioecological model of human development defines the home environment as a key context for proximal processes that influence childhood development (Bronfenbrenner, 1986; Bronfenbrenner & Ceci, 1994; Bronfenbrenner & Morris, 2006). In particular, Bronfenbrenner argued that stability in home environments was particularly important for healthy development (Bronfenbrenner & Evans, 2000), and disruptions to routine family life were centrally important in poor childhood psychological functioning (Bronfenbrenner & Ceci, 1994).

Identifying the features of stable and consistent family home environments, and examining the effects of variability in home contexts, has therefore been important for testing the propositions of the bioecological theory of development. To achieve this aim, however, home environmental features must be recorded or measured in some way. The purpose of the current study, therefore, was to examine the measurement properties of a widely-used short-form scale, which was developed with the intention of capturing variability in home environments, the Confusion, Hubbub And Order Scale (CHAOS; Matheny et al., 1995). We begin with some background on the measurement of home environments before providing a brief history of the CHAOS measure.

Capturing the Variability in Home Environments

In the mid-20th century researchers recorded information about differences between home environments via in-person observations (e.g. Wilson & Matheny, 1983). Research assistants spent many thousands of hours attempting to unobtrusively observe aspects of home environments, including interactions between parents and chil-

[a] Corresponding Author: Sally A. Larsen, School of Education, University of New England, Armidale, NSW, 2351. Email: slarsen3@une.edu.au

dren, the number of visitors coming and going, ambient noise levels (both internal and external to the home), and observable routines established by parents, among other features (Evans, 2006). These efforts to measure the quality of home environments were expensive, and limited in that a finite number of households in a geographically constrained area could be visited by research assistants within a given timeframe. In the 1980s, therefore, measurement of household environments began to shift from a reliance on observations, recorded as both qualitative information and observer ratings on quantitative scales, to self-report scales, where parents were asked to rate aspects of their homes according to questionnaire items.

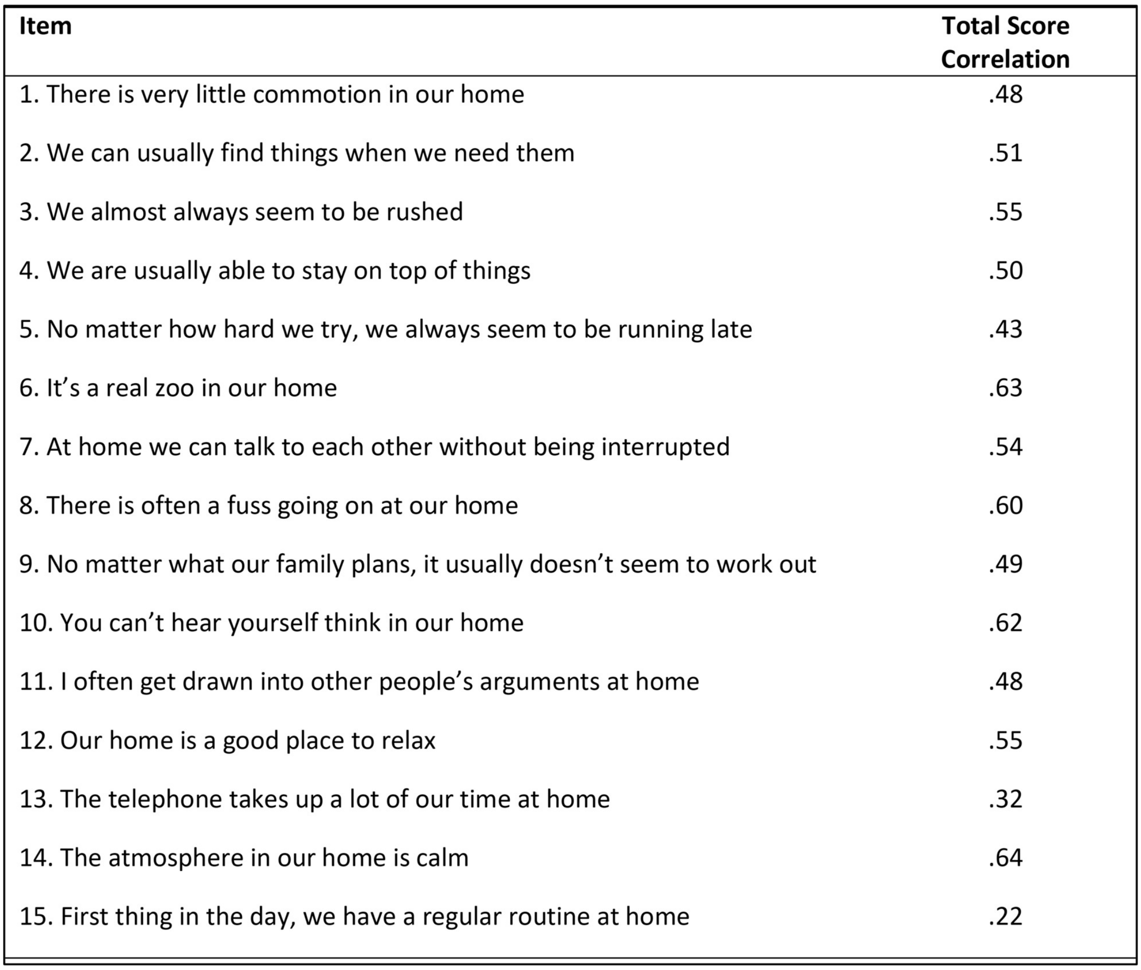
The CHAOS measure was one of several self-report instruments established during the latter decades of the 20th century (the HOME scale is another widely used instrument; c.f. Bradley, 2015 for a review). At least two versions of the CHAOS scale have been used in research since its inception: the original 15-item version proposed by Matheny et al. (1995), and a short-form 6-item version. The scale has been used extensively in research: there are over 500 citations to date of the paper reporting the psychometric properties of the original scale (Matheny et al., 1995), and a library database search identified 305 articles and 62 dissertations referencing the name of the scale (June, 2022). It is difficult to identify how many of these research articles use the long form and how many use the short form of the scale given that both have the same title. The latter, short-form is the focus of this study, however some background on the original long-form is relevant here.

The 15-item CHAOS measure was developed in a sample of over 400 families of twins participating in the Louisville Twin Study during the 1980s. The original scale (reproduced in Figure 1.) comprised true/false scored items, half reverse scored, which were summed to produce an overall measure of the (in)stability of the household environment. Conceptually the scale captured various aspects of household confusion and disorder, including high levels of noise, clutter, disorganization and “frenetic activities” (p.432). Matheny et al. (1995) validated the 15-item scale using a subsample of 123 mothers of infants ranging in age from 6 to 30 months, reporting a reliability coefficient of α = .79. A further subsample of 42 mothers completed the questionnaire at a 12-month interval with a test-retest correlation of r =.74. Matheny et al. noted that the 15-item CHAOS scale accounted for a unique proportion of systematic within-home differences that could not be attributed to parent education or SES measures. Nonetheless, direct observations of the home environment were not interchangeable with the CHAOS scale: there was a significant, but not high, degree of overlap between the two measures in the study (R^2 = .39). The 15-item scale was further validated in two different samples of preschool (n = 106) and school-aged children (n = 676; Dumas et al., 2005), demonstrating overlap with, but distinction from, measures of socioeconomic status, and good internal consistency reliability (α = .83 / .81).

The short 6-item form of the CHAOS measure was first used in the late 1990s by studies including the Twins Early Development Study (TEDS; e.g. Asbury et al., 2003) and the Western Reserve Reading and Math Project (WRRMP; e.g. Hart et al., 2007). The short-form of the scale is reproduced in Figure 2. This form consists of 4 items from the original Matheny et al. (1995) 15-item scale, plus two items not appearing in the original scale: Item 1 *The children have a regular bedtime routine* and item 5 *There is usually a television turned on somewhere in our home*. There is no published information on why these 6 items were chosen for inclusion in the short-form CHAOS and no evaluations of whether the short form captures the full range of the original intended construct (e.g. Smith et al., 2000). In published articles using the short-form, construct validity evidence is universally attributed to the article which reports the validity of the long-form, 15-item scale (Matheny et al., 1995), and reliability information is usually reported as Cronbach’s coefficient alpha, with estimates ranging from α = 0.52 - 0.68 (see Supplementary Table S1).

Notwithstanding the lack of published evidence that the short-form CHAOS scale reliably measures the same construct as that proposed by the long-form, many studies have used the 6-item scale to examine links between the home environment and childhood functioning. Table S1 in the supplementary material shows details of 22 papers published between 2003 and 2023 that used the short-form CHAOS. We acknowledge these papers are a snapshot of all studies that have used this scale: in a scoping review of the relationships between CHAOS and child, parent and family outcomes, Marsh et al. (2020) identified 42 studies that had used the short-form CHAOS up to 2018. Published research has examined associations between CHAOS and cognitive development (Petrill et al., 2004; Pike et al., 2006), reading skills (Johnson et al., 2008), language development (Asbury et al., 2005), and behavioural problems (Coldwell et al., 2006; Deater-Deckard et al., 2009; Laurent et al., 2014; Peviani et al., 2019), consistently finding that higher ratings of household CHAOS are associated with poorer functioning or development. The short scale has also been used in studies examining social determinants of health and wellbeing (Ganasegeran et al., 2017; Suku et al., 2019). Many of the projects that have collected data on the CHAOS short-form have been twin studies which examine home environments within the behaviour genetics theoretical framework: home environments are considered an aspect of *shared environments,* i.e. environmental features which serve to make twins more similar to one another (Plomin et al., 2013), although it is acknowledged that twins can perceive the same objective environment differently (Hanscombe et al., 2010). Several papers have reported the extent to which home environments mediate or moderate genetic influences on childhood outcomes (Asbury et al., 2003; Gould et al., 2018; Harlaar et al., 2005; Hart et al., 2007; Petrill et al., 2004; von Stumm et al., 2023), and one attempted unsuccessfully to identify genetic influences on reports of CHAOS using a genome-wide association design (Butcher & Plomin, 2008).

The short-form scale has also undergone several additional transformations, including translations into languages other than English (e.g. Deater-Deckard et al., 2019;



Item	Total Score Correlation
1. There is very little commotion in our home	.48
2. We can usually find things when we need them	.51
3. We almost always seem to be rushed	.55
4. We are usually able to stay on top of things	.50
5. No matter how hard we try, we always seem to be running late	.43
6. It's a real zoo in our home	.63
7. At home we can talk to each other without being interrupted	.54
8. There is often a fuss going on at our home	.60
9. No matter what our family plans, it usually doesn't seem to work out	.49
10. You can't hear yourself think in our home	.62
11. I often get drawn into other people's arguments at home	.48
12. Our home is a good place to relax	.55
13. The telephone takes up a lot of our time at home	.32
14. The atmosphere in our home is calm	.64
15. First thing in the day, we have a regular routine at home	.22

Figure 1. Original 15-item Confusion, Hubbub and Order Scale (CHAOS) from Matheny et al. (1995)

Note. **Parents responded true / false to each item. Scores were summed to create a composite with higher scores representing greater 'chaos'.**

Ganasegeran et al., 2017), and versions where children or adolescents themselves rate their home environments on a three-point likert response scale. Using this adolescent self-report data, studies have examined links between CHAOS and academic achievement, behavioural functioning (Hanscombe et al., 2010, 2011; Kim-Spoon et al., 2017), and brain activity in functional MRI studies (Lauharatanahirun et al., 2018). An even shorter version of the short-form, using data from only five items, is also evident in the literature. This variation is used in the Parenting across Cultures Study, which recruited 511 urban families in six low-to-middle income countries (China, Kenya, the Philippines, Thailand, Colombia and Jordan; Deater-Deckard et al., 2019). While responses on all six items were initially collected, item 5 *There is usually a television turned on somewhere in our home,* was subsequently omitted due to poor face validity: it is logical that families in low income countries may not own televisions. Studies emerging from the Parenting Across Cultures project also report the original Matheny et al. (1995) article as evidence of the reliability of the scale (Chang, Lu, Lansford, Bornstein, et al., 2019, p. 4; Chang, Lu, Lansford, Skinner, et al., 2019; Deater-Deckard et al., 2019).

Given the widespread use of the short-form CHAOS, and its attractiveness in terms of minimal time commitment of respondents in multivariate surveys, the lack of reliability and construct validity information for the scale is of concern. The present study aimed to examine the measurement properties of the short-form CHAOS scale. We used several approaches and five datasets to adjudicate whether we could conclude that the scale is valid and reliable for measuring the quality of home environments. In defining validity, we take the position of Borsboom et al. (2004) who argued simply that "a test is valid if it measures what it purports to measure" (p.1061). Furthermore, we define reliability as "an index of measurement precision" (p. 1070) that can be evaluated within a scale (i.e. how well do items measure the same construct) and across measurement occasions (i.e. between samples or within samples over time). In this study we therefore: 1) examine the factor structure of the six items, 2) evaluate whether the measure is invariant across groups, 3) evaluate whether the measure is invariant within groups over time, and 4) examine the predictive validity of the scale using socioeconomic status and childhood academic achievement as criterion variables. In this way we collate evidence of the validity and reliability of the scale to measure the quality of home environments in different populations.

Figure 2. Six items in the short-form version of the CHAOS scale with variations for different studies

Instructions: Below are some things that happen in most homes. Please circle the number that best describes your home: *Response options:* (1) Definitely untrue / (2) Somewhat untrue / (3) Not really true or untrue / (4) Somewhat true / (5) Definitely True	Abbreviation
Items	1. BedRoutine [i]
1. The children / the twins / my child have (has) a regular bedtime routine* (e.g., same bedtime each night, brushing teeth, reading a story/book) (ADSAT, FTP-RBE, Project KIDS) 1. The twins / the children have a regular bedtime routine* (for example, same bed each night, a bath before bed, reading a story, saying prayers) (TEDS; WRRMP)	
2. You can't hear yourself think in our home	2. HomeNoise
3. It's a real zoo in our home	3. HomeZoo
4. We are usually able to stay on top of things	4. HomeControl [i]
5. There is usually a television turned on somewhere in our home*	5. HomeTV
6. The atmosphere in our house is calm	6. HomeCalm [i]

Note. * indicates item that did not appear in the original 15-item scale. [i] Indicates variables reverse-coded for analysis so that higher scores = greater 'chaos'.

Measurement Invariance

Combining multiple survey items into a single composite score is very common practice in social science research. Creating a sum or average from several items, however, assumes that the scale in question captures one underlying factor (McNeish & Wolf, 2020). In cases where the construct of interest has been shown to capture a single factor, using composites is a defensible strategy (Widaman & Revelle, 2022), however, there is minimal documentation regarding the most appropriate factor structure of the short-form CHAOS (except in Johnson et al., 2008, which suggested two factors). Furthermore, use of a scale in different populations or in longitudinal designs assumes that the measure captures the same latent construct regardless of context or measurement occasion (Millsap & Olivera-Aguilar, 2012). Any differences in the means or variances of the observed items, either between groups or within groups over repeated measures, is assumed to be related to differences in the latent construct itself, rather than exogenous differences that are unrelated to household order and routine.

In this study therefore, we aimed to test the assumptions that underlie the common usage of the CHAOS short-form. These assumptions include a) that a single factor underlies the six items in the short-form CHAOS, b) that the factor structure is the same across samples, c) that the factor structure is invariant over time, and d) that differences in the observed variables are caused by differences in the latent construct, and are not due to unobserved differences in response patterns unrelated to the household environment.

We can begin to test these assumptions about the CHAOS measure using existing datasets and employing confirmatory factor analyses, with restrictions to test for measurement invariance. We follow the typical procedure for testing between-group measurement invariance recommended by methodologists (e.g. Millsap & Olivera-Aguilar, 2012; van de Schoot et al., 2012): namely, the same confirmatory factor model is first estimated in each group, then increasingly restrictive equality conditions are introduced for different sets of parameters. Longitudinal invariance testing proceeds in the same manner, though rather than differences between groups, differences in repeated measures within groups is examined. If measurement invariance holds across samples, we can be confident that comparing the results of studies using the CHAOS scale in different contexts is valid and informative. If longitudinal measurement invariance holds within groups, we can be confident that differences in means over time are due to changes in the latent measure of household order and routine, rather than differences in response patterns unrelated to the construct of interest. On the other hand, if the analyses indicate that the measure is non-invariant – either between groups, or over time – response patterns on the observed items could be systematically influenced by unobserved, exogenous factors, for example interpretive differences for specific items, rather than differences in the latent domain of interest.

This Study

Our hypotheses for the study are informed by, 1) the consistently low reliability reported in studies using the short-form CHAOS (α = .52 - .68), 2) evidence from one study that the six items are better represented by two factors rather than one (Johnson et al., 2008), and 3) preliminary evidence generated by an exploratory factor analysis of the six items indicating a two-factor solution (see below). Given this information, we hypothesised that a two-factor dimensional structure would better fit the data on the CHAOS items in all samples. Preliminary analyses also informed our hypothesis that the measure would be non-invariant across the five samples: that is, we did not expect the six items to behave similarly within all samples, nor did we expect the factor structure to be repeatable across samples. The only information on the longitudinal stability of the CHAOS measure comes from test-retest correlations of the sum score of six items (e.g. Deater-Deckard et al., 2009; Laurent et al., 2014; Matheny et al., 1995; Petrill et al., 2004). Accordingly, any attempt at a precise hypothesis on whether the measure will be longitudinally invariant will be somewhat limited. Nonetheless, we registered a hypothesis of longitudinal measurement non-invariance since we had no evidence to suggest the scale would operate more consistently over time within samples, compared with between groups. Finally, and based on the research findings described above, we predicted that higher CHAOS would be negatively associated with both family socioeconomic status and academic achievement, however in the case where a two-factor model is most appropriate, it was not clear whether one or both factors would be significantly associated with each criterion variable. Johnson et al. (2008) demonstrated that only one of two factors (*household order* but not *noise*) was associated with several measures of childhood literacy, however whether the factor structure identified in this previous analysis holds across all samples will only become evident after the initial between-group invariance testing.

Ethical approval for this study was obtained from the first author's institution (Human Research Ethics Committee Approval# HE22-093). Preliminary hypotheses and an overview of the study were preregistered at the Open Science Framework (https://osf.io/akmf4). Subsequent to preregistration we gained access to an additional dataset not noted in the preregistration (the Project KIDS data). We note also that our preregistered analysis plan was not specific about the preliminary analyses we would perform. We elaborate here for completeness. Since the first author had access to the data collected in the Academic Development Study of Australian Twins (ADSAT), an initial exploratory factor analysis (EFA) had already been carried out using data from the first wave of collection (n=596). We report this EFA below. Due to the longitudinal nature of participant recruitment in the ADSAT, the data used for the confirmatory analyses in the current study came from the participants recruited subsequent to the initial wave (n = 1294).

Our preliminary tests of measurement invariance included data from only the ADSAT and Florida Twin Project on Reading, Behavior and Environment (FTE-RBE). The inability to confirm measurement invariance across these two datasets informed our preregistered research plan, and indeed prompted the expansion of the study to include multiple data sources. Adding to the complexity, we only subsequently identified that the initial EFA undertaken with the ADSAT subsample suggested a different configuration of items to factors compared to that shown in Johnson et al. (2008), noted above. For this reason, we opted to test both two-factor configurations in an attempt to identify whether one model was a better fit to the data in multiple samples than the other.

Methods

Secondary data for the project was sourced from: three studies located in the US, the Western Reserve Reading and Math Project (WRRMP; Hart et al., 2007; Petrill et al., 2006), the Florida Twin Project on Reading, Behavior and Environment (FTP-RBE; Taylor et al., 2019), and Project KIDS (Kids and Individual Differences in Schools; van Dijk et al., 2022); one study located in the UK, the Twins Early Development Study (TEDS; Oliver & Plomin, 2007; Rimfeld et al., 2019); and one study located in Australia, the Academic Development Study of Australian Twins (ADSAT; Larsen et al., 2020). These studies were selected because all collected parent reports on the English language short-form CHAOS using a five-point likert response scale (see Figure 2), and the children of interest were aged between 3 (earliest wave of TEDS) to 12 years (upper age of FTP-RBE, ADSAT and Project KIDS wave 1 samples). Descriptive statistics of participants in all samples and data collection waves are in Table 1.

Samples and Measures

The *Academic Development Study of Australian Twins (ADSAT)* recruited a national sample of 2762 families of Australian school aged twins between 2012 and 2017 (Larsen et al., 2020). The design of study recruitment was partly prospective and partly retrospective. For the main analysis in the current study we selected families who were recruited and had completed the CHAOS measure when their twins were in Grades 3, 4 or 5 (n=1294; age 8 to 11 years). This age group was selected in an attempt to align the ages of participating children as closely as possible across samples. For the initial exploratory factor analyses we used an additional sub-sample of n=596 families participating in the ADSAT who also completed the CHAOS form on enrolment into the study. This subsample was not included in subsequent confirmatory models. Parents completed the CHAOS measure only once.

In the ADSAT, academic achievement was measured by standardized scores on reading comprehension and mathematics tests undertaken by children as part of the National Assessment Program: Literacy and Numeracy (NAPLAN; Australian Curriculum Assessment and Reporting Authority, 2017). For this study we used scores on the Grade 3 assessments to align with when the CHAOS items were collected. Socioeconomic status in this dataset is a factor score comprising the highest educational attainment of both parents, the occupational prestige ranking of both parents and an indicator of neighbourhood socioeconomic advantage (see Larsen et al., 2020 for details). While the academic achievement of students participating in the ADSAT was representative of the locations that participants were drawn from (largely metropolitan areas of Australia), we note that the parents of study participants were generally more highly educated than the Australian population and 95% of participants indicated Australian, UK or other European ancestry (Larsen et al., 2020). The study also did not include participants from the Northern Territory region of Australia, which comprises the largest proportion of the most disadvantaged Indigenous and Torres Strait Islander peoples of all Australian states and territories.

The *Florida Twin Project on Reading, Behavior and Environment (FTP-RBE)* is a subsample of the 2753 twin pairs recorded in the Florida State Twin Registry (FSTR; Taylor et al., 2019). Beginning in 2012, a subsample of families with twins enrolled in the FSTR were invited to enrol in the FTP-RBE, which involved completing a questionnaire, containing in part the CHAOS items, every other year over six

Table 1. Descriptive statistics of the five samples included in the analysis

Study Sample (Acronym)	Country	Wave	*N*[i]	Female [ii] (%)	Age [iii]		α [iv]	ω_h [v]
					M	SD		
Academic Development Study of Australian Twins (ADSAT)	Australia	1	1294	50%	8.79	0.45	0.67	0.54
Florida Twin Project on Reading, Behavior and Environment (FTP-RBE)	USA	1	568	46%	11.16	2.52	0.55	0.37
		2	437		13.30	2.44	0.63	0.53
		3	313		15.24	2.51	0.50	0.48
Project KIDS	USA	1	442	49%	11.07	3.07	0.59	0.50
Western Reserve Reading and Math Project (WRRMP)	USA	1	580	57%	6.09	0.69	0.68	0.56
		2	512		7.16	0.67	0.65	0.29
		3	494		8.21	0.82	0.70	0.63
		4	352		9.81	0.98	0.62	0.58
		5	362		10.90	1.01	0.67	0.37
		6	368		12.21	1.20	0.64	0.45
		7	246		15.05	1.45	0.59	0.48
Twins Early Development Study (TEDS)	UK	3[vi]	6009	50%	3.01	0.14	0.63	0.44
		4	8014		4.03	0.15	0.66	0.59

Note. [i] *N*=families; for twin studies the number of twins is twice the number of families. [ii] Proportion as at study commencement. [iii] Age calculated in years: decimal places indicate proportion of a year. [iv] Cronbach's Alpha calculated for all six items. [v] McDonald's omega (hierarchical). [vi] Waves 1 and 2 of TEDS did not include the CHAOS items.

years (i.e., three waves of questionnaire assessment). The mean age of twins for the first wave of the questionnaire data collection was 11.16 years. In total, 568 families (72% of the invited participants) provided data on the CHAOS at wave 1, reducing to 447 at wave 2 and 313 at wave 3. Academic achievement was measured by scores on the Florida Comprehensive Assessment Tests (FCAT) reading subtest, undertaken by students in the 2011-12 and 2012-13 school years. The FCAT is a standardized assessment of reading, completed by students at the end of grades 3 to 11. FCAT data were provided by Florida's Progress Monitoring and Reporting Network (PMRN). Socioeconomic status is a factor score generated using five observed variables: estimated family income, both parents' highest educational attainment, and both parents' occupational prestige. The initial recruited sample comprised 49% white, 19% African American, and 24% Hispanic background students, with 60% eligible for free or reduced-price lunch (Little et al., 2019).

Project KIDS is a repository of data collected in nine randomized control trials of reading interventions undertaken in the North Florida region of the United States between 2005 and 2011 (see van Dijk et al., 2022). Data on the CHAOS short form was collected in 2013 from a subsample of 442 families of singleton children who had participated in at least one trial. Data on school achievement was collected in the same parent survey. For both English Language Arts and Math, parents reported their children's achievement on a 5-point rating scale, ranging from *A/Excellent* (1) to *F/Fail* (5). Achievement variables were reverse coded before analysis so that higher ratings indicated better achievement, similar to other achievement tests used in this study. Socioeconomic status variables and factor score estimation was identical to that in the FTP-RBE study described above. The sample comprised mostly white (54%) and African American (34%) background students, with 30% indicating eligibility for free or reduced-price lunch (though 34% of participants had missing data on this indicator; van Dijk et al., 2022).

Western Reserve Reading and Math Project (WRRMP) is a longitudinal cohort-sequential study which recruited families of twins, primarily in the state of Ohio, USA, beginning in 2002 (Petrill et al., 2006). Twins were in kindergarten or first grade on recruitment (mean age = 6.09 years). Measures of literacy and CHAOS were collected approximately annually over seven waves of data collection. Across all waves, 794 families provided at least some data to the project. The short-form CHAOS was collected at each wave of the study, with 580 families answering the items in wave 1, reducing gradually to 246 families responding by wave 7.

We selected five assessments of academic skills in both reading and math domains collected across all waves of the WRRMP. These included, a) two assessments of reading comprehension, the PIAT-R/NU (Dunn & Markwardt, 1998) and the WRMT-R passage comprehension assessments (Woodcock, 1987), and b) three assessments of math subdomain skills, the Woodcock-Johnson calculation, applied problems, and quantitative concepts tests (Woodcock, 1987). All children who were able to be followed up at each wave provided data on these assessments. For the WRRMP study we used a proxy of socioeconomic status using variables that were available in the dataset: an average of both parents' highest educational attainment. Most parents reported some college study (39%), bachelor's degree (30%) or some postgraduate education (25%), with only 1-2% attaining high school or less. The majority of the sample were white (92%) and came from two-parent households at the time of recruitment (94%; Petrill et al., 2006).

The *Twins Early Development Study (TEDS)* recruited a nationally representative sample of 13,732 families of infant twin pairs in the United Kingdom from 1994-1996 (Oliver & Plomin, 2007). For this study we use wave 3 and 4 of data collection, when twins were aged 3 and 4 years, respectively. Parents responded to the short-form CHAOS in both waves, with 6009 parents providing data in wave 3, and 8014 in wave 4. Later collections on the CHAOS measure used a 3-point likert response scale and/or asked twins themselves to respond, rather than parents. We omit these waves and focus on the CHAOS measure obtained in the same manner as that for the other data collections in this study. Academic achievement was assessed when twins were aged approximately 7 years. All students in the UK undertake National Curriculum assessments in core subjects. Standardized assessment results for English and Mathematics were sourced from government data collections. We use the socioeconomic status variable available in the TEDS dataset, a composite generated from five variables: occupational prestige of both parents, highest educational levels of both parents, and mother's age at the birth of the first child. The TEDS sample was representative of the racial composition of the UK population (93% white in both sample and population), though parents were more highly educated on average: 40% had obtained A-levels or higher by wave 4 (twins aged 4 years), compared with 32% of the UK population (Oliver & Plomin, 2007).

We note that four of the five studies included in this project were studies of child and adolescent twins. For each dataset, the CHAOS items and SES variables were collected at the family level (i.e. one set of responses by family for each wave in each study), therefore we did not need to account for the nested nature of data collected on twin pairs. Academic achievement variables were collected for each twin separately, however, so in instances where we use achievement as criterion variables, we selected one twin from each pair. We do not report results for the second twin, but findings were no different.

Analysis Plan

In this study we aimed to test the factor structure, between-group and longitudinal measurement invariance, and predictive validity of the short-form CHAOS using five samples collected in different contexts. We first wanted to test whether the usual approach to using the six items – i.e. combining them into a single mean or sum score – is the optimal approach to the use of the scale. Only one study to date has reported an exploratory factor analysis (EFA) of the items (Johnson et al., 2008). Using the WRRMP Wave 1 data this study demonstrated a two-factor solution in an EFA with one factor comprising items 1, 4 and 5 (termed "household order and routine"), the second factor comprising items 2, 3 and 6 ("quietness of the household"; Johnson et al., p. 5). The two factors correlated at r = .33. The proportion of variance explained by the two-factor solution and the factor loadings of the items were not reported. Items were subsequently summed within each factor for further analyses. Given this study is the only one to date to examine the factor structure of the short-form CHAOS, the first step in the analysis for the current study was an EFA using a separate subsample of participants in the ADSAT (n=596) that were not subsequently included in the confirmatory factor models. Specifically, a principal components analysis using full information maximum likelihood estimation was undertaken. Due to the results in Johnson et al. we expected that a two-factor solution would fit the data better than a one-factor model. Therefore, we examined eigenvalues, compared the proportion of variance explained by one- and two-factor solutions, and examined item-factor loadings.

To further examine whether a one- or two-factor structure of the six items was best supported by all datasets, we next ran confirmatory factor analyses (CFA) separately for each sample. We standardized the latent variables to ensure that models were identified and to allow examination of differences in item loadings and intercepts. For CFAs comprising two factors we allowed factors to correlate, but did not allow any cross-loadings of items, nor any residual covariances. We examined model fit statistics, and compared nested models to identify the best solution in each sample. We predicted that two-factor models would be a better fit to the data than one-factor models for all samples, however we made no specific predictions about whether the configuration of items reported by Johnson et al. (2008) would be the best fit in each sample.

Next, measurement invariance was examined via multiple-group confirmatory factor models. In this step we consider each dataset a different group since each study was conducted in a different context, and three countries are represented by the five datasets, Australia, the UK and the USA. We followed the procedure suggested in several sources and tested i) configural, ii) metric, and iii) scalar invariance (e.g. Byrne, 2012; Meredith & Teresi, 2006; Millsap & Olivera-Aguilar, 2012; Putnick & Bornstein, 2016). We did not expect strict invariance (i.e. invariance of residuals) to hold across groups so planned to test this step only where scalar invariance was confirmed. Specifically, configural invariance models force the same factor structure across groups but allow item loadings, item intercepts and residuals to vary. Because we planned to first test confirmatory factor models for each group, and select the best-fitting model, we expected configural invariance to hold. Metric invariance forces equivalence of factor loadings across groups and assesses whether this restriction leads to a significant reduction in model fit. Scalar invariance tests for equivalence of item intercepts across groups retaining the equivalence of factor loadings tested in the previous step. Strict invariance retains the equivalence constraints introduced by metric and scalar invariance, and constrains item residuals to equality. If at any step model fit statistics suggested significantly poorer fit, we examined the parameters constrained by that step to identify potential sources of model misfit.

We next aimed to assess whether the CHAOS was longitudinally invariant– i.e. measuring the same construct over time – given the most defensible factor structure of the items. We tested longitudinal invariance in three of the samples that contained multiple waves of data – TEDS, FTP-RBE and WRRMP. Longitudinal invariance proceeds in the same set of steps as multiple-group invariance testing, described above, however rather than comparing across groups, measurement is compared across waves of data collection.

Model fit for CFAs and invariance testing models was assessed using several fit statistics. Given that χ^2 goodness of fit is affected by large samples or variable sample sizes in multiple group models (Byrne, 2012), we report this statistic along with several others. In particular, we examine the root mean square error of approximation (RMSEA), which ideally should fall $\leqslant$ 0.08, and the Standardized Root Mean Square Residual (SRMR) which ideally should be $\leqslant$0.05 (Byrne, 2012). We also examine the comparative fit index (CFI), which provides an estimate of incremental fit of the model compared with a baseline model. Current advice suggests CFI values of $\geqslant$ 0.95 indicate adequate model fit (West et al., 2012).

For assessing the model fit of the nested models, such as those in each step of the measurement invariance tests, we examine the change in χ^2 relative to change in degrees of freedom (*df*). Ideally the change in χ^2 for each *df* should have p >.001, indicating that the more restricted model is not a worse fit to the data than the less restricted model. When equating parameters across groups in measurement invariance analyses, particularly when large numbers of groups are compared, ΔRMSEA and ΔCFI can also be examined (Rutkowski & Svetina, 2014). A change of 0.010 (RMSEA) and -0.010 (CFI) are indicative of non-invariance between groups when parameters are constrained to equality for metric or scalar invariance tests (2002; OECD, 2010). Finally, Akaike's Information Criterion (AIC) can provide additional information about fit for non-nested models, with smaller values indicating better model fit. We report and interpret ΔAIC where appropriate (West et al., 2012).

It is important to note that interpreting change in model fit statistics to assess measurement invariance across more than two groups, as we do in this study, can generate information without clear or simple interpretations. For example, should model fit decrease significantly at any step of measurement invariance testing, with five groups in the model, it may not be clear whether one sample is driving model misfit, while others are sufficiently comparable. Notwithstanding this interpretational problem, the main aim of the study is to evaluate whether the CHAOS measure behaves similarly across contexts, therefore non-invariance of even one sample is problematic for the applicability, use, and interpretation of the scale in different contexts.

Finally, we planned to examine the predictive validity of the CHAOS measure using two criteria. We examined zero-order correlations between CHAOS and a socioeconomic status variable (or proxy), and the academic achievement variables available in each dataset. We compared results using a) a one-factor model of CHAOS, b) a two-factor model, and c) analyses where the CHAOS items are composed as factor scores, with results when items are composed as mean scores, as is more common in the published literature. Because there are multiple tests for each dataset we use a conservative p-value cut-off of p <.001 to denote statistically significant correlation coefficients.

All analyses were run in the statistical program R (R Core Team, 2020) using the psych package (Revelle, 2022) for descriptive statistics, reliability statistics, creating factor scores and exploratory factor analyses, the lavaan package (Rosseel, 2012) for confirmatory factor models and invariance testing, and ggplot2 (Wickham, 2014) for figures. Code for confirmatory factor analyses, and invariance testing is at the OSF (https://osf.io/akmf4). Data from FTP-RBE, Project KIDS and the WRRMP is available at LDBase repository (Hart et al., 2020). Data from the ADSAT is available on request to the first author, and data from TEDS is available on request from data managers (Kings College London, 2022, https://www.teds.ac.uk/researchers/teds-data-access-policy).

Results

For each sample, means, standard deviations, skew, and kurtosis of each item, and zero-order correlations between items were generated. These are reported in Tables S2-S12 in the supplementary material. We report all waves of data for multi-wave studies, except WRRMP, where we report waves 1 through 4. Remaining waves of WRRMP showed similar correlation patterns. Correlations between items were all positive, with some variation in the strength of correlations across the samples. Perhaps most notable were the differences in correlations between item 5. *There is usually a television on somewhere in our home*, and the remaining items. In the Project KIDS and FTP-RBE samples, correlations between this item and the remaining five were generally smaller ($r \leqslant$.17) than those in the ADSAT, WRRMP and TEDS samples ($r \leqslant .34$). On the other hand, the strongest correlation in all samples was between items 2. *You can't hear yourself think in our home* (reversed) and 3. *It's a real zoo in our home* (r = .56 - .77).

Variation was also evident in item means and distributions across studies. Figures S1-S5 (supplementary materials) show item distributions for each dataset selecting one wave from each multi-wave study. It is notable that two of the item distributions were positively skewed in all samples (item 1. *The children have a regular bedtime routine,* and item 4. *We are usually able to stay on top of things*). Item 3 *It's a real zoo in our home* was positively skewed in some samples (FTB-RBE and Project KIDS) but not others. Item means also differed between samples for some items. For example, the mean of item 2. *You can't hear yourself think in our home* (reversed) varied from 1.95 (Project KIDS) to 3.29 / 3.27 (TEDS sample, wave 1 / wave 2). Similarly, means for item 5, *There is usually a television on somewhere in our home,* and item 3 *It's a real zoo in our home* differed by sample. By contrast, response patterns within studies (i.e. where respondents answered the CHAOS several times over data waves) were comparable.

Exploratory factor analysis. To provide additional information about whether a one-factor or a two-factor solution would be most appropriate across all samples, we undertook an exploratory factor analysis (EFA) on a subsample of participants in the ADSAT (n=596; 47% female). We selected the ADSAT data because the first author had access to these data before obtaining permission to access the remaining four datasets. Table S13 (in the supplementary materials) shows eigenvalues and proportion of variance explained for each principal component in the EFA, estimated using the maximum likelihood method. The first two components collectively explained 57% of the variance, and both had eigenvalues > 1. Remaining components had eigenvalues < 1, and the parallel analysis plot (Figure S6 in supplementary materials) also supported a two-factor solution.

Factor loadings and communalities for a two-factor solution are reported in Table S13 (supplementary material). Interestingly the EFA in this sample suggested a different pattern of items to factors than that indicated by the published example using the WRRMP data (Johnson et al., 2008). In the WRRMP data, items 1, 4, and 5 comprised one factor, termed '*order and routine*', and items 2, 3 and 6 comprised the second factor, labelled '*quietness of the household*'. In the ADSAT data, by contrast, items 2, 3 and 5 loaded on one factor and appeared to represent *household noise*, while item 6 cross-loaded on both factors. This cross-loading suggests the wording of item 6, *The atmosphere in our house is calm*, could be interpreted in the light of either household noise, or routine. Given this inconsistent result, we opted to test two different configurations of the six items in confirmatory factor models. The configuration identified by Johnson et al. grouped items 1, 4 and 5 (*routine*), and 2, 3 and 6 (*quietness*). The second configuration informed by the EFA described here grouped items 1, 4 and 6 (*disorder*), and items 2, 3, and 5 (*noise*).

Confirmatory Factor Analysis by sample. To evaluate whether the one-factor, or either of the proposed two-factor structures of the six items was consistently reproduced over the five samples, we first tested the three models separately in all samples and waves. First, and in alignment with the common usage of the scale as a sum or average of the six items, we tested a one-factor model, forcing all items to load on one latent variable with no residual correlations (Table 2, Model A. in all samples and waves). We compared this one-factor model with the two different configurations of a two-factor model, each allowing three items to load on each factor (see justification above). Because the two-factor models are not nested (the same number of parameters is estimated in both) we compare each (i.e. Models B. and C. in each panel of Table 2.) with the one-factor model. While this comparative process is imperfect given that Models B. and C. cannot be directly compared to each other using most fit statistics, evaluating the fit of each model against the one-factor option does provide some information about which solution may be more appropriate. In addition, we interpret the AIC for additional information about which two-factor model might be retained.

In all samples, a one factor model (A.) was a poor fit to the data according to all criteria (Table 2., first row of each panel). In all cases the RMSEA statistic did not fall within the acceptable range, and the CFI and TLI statistics were <.95. Model B tested the two-factor solution reported in Johnson et al. (2008), with factors termed '*quietness*' and '*routine*'. Change in χ^2 (*df*), RMSEA, CFI and AIC for the two-factor model compared with the one-factor model showed an improvement in fit in all samples. Nonetheless, in most cases fit statistics were poor or borderline. The exception was the WRRMP dataset (that used by Johnson et al.), which showed borderline-good model fit for this configuration of items in five of seven waves.

Model C tested the alternative configuration of the six items suggested by the EFA in the subsample of the ADSAT data. Model C fit the data better than the one-factor model according to all fit statistics (Table 2, model C.). AIC statistics indicated that this alternative two-factor configuration was a better fit to the data than that tested in model B for all samples and waves except for the WRRMP data. In the ADSAT, Project KIDS, wave 2 of the WRRMP, and both waves of TEDS data, fit statistics were acceptable or borderline for model C. The fit of model C was acceptable for wave 2 of the FTP-RBE data, however none of the models demonstrated adequate fit for waves 1 or 3 when evaluating the RMSEA, CFI and TLI against suggested cut-off criteria. In particular, model C for wave 3 returned Heywood cases (i.e. negative variances) for two of the observed variables. Similarly, the fit of all of the models in wave 3 of the WRRMP remained poor.

Notwithstanding these problems, in order to assess between-group measurement invariance, we selected the following data waves from multiple wave studies: We retained wave 1 of the FTP-RBE data because this wave had the least missingness; we retained wave 2 of the TEDS sample (older age group and least missingness), and wave 2 of the WRRMP data (best fit for model C.). We included also the ADSAT confirmatory sample and the complete Project KIDS sample. Selecting one wave from multiple-wave studies is problematic, however, if the measure is longitudinally invariant the results should not differ depending on data collection wave. We report results for longitudinal invariance following results for between-group invariance.

Measurement Invariance. Table 3 shows fit statistics for each step of invariance testing incorporating all five samples. The configural invariance model forces the same configuration of items loading on factors across all groups with no cross-loadings or residual covariances for observed items. Factor loadings, intercepts, variances and covariances are allowed to vary by group. Notwithstanding the poor fit of the models for some individual samples noted above, the configural invariance model (Table 3, Model 1) showed borderline acceptable fit to the data when evaluated by the RMSEA (0.077, 90% CI [0.071, 0.083]), CFI (0.957), and SRMR (0.035) statistics. Next, model 2A. (Table 3) tested for metric invariance by constraining factor loadings of all items to equivalence across groups. According to the AIC and the χ^2 difference relative to degrees of freedom ($\Delta\chi^2$ (Δdf) = 99.10 (16), *p* <.001), the fit of model 2A was significantly worse than model 1. However, the ΔRMSEA (0.007), and ΔCFI (0.007) indicated the fit of this model was not worse relative to model 1 (using cutoff values of 0.010 for each; Rutkowski & Svetina, 2014). Given this mixed information, we examined the factor loadings across the five samples in the configural invariance model. The load-

Table 2. Model fit statistics testing one- and two-factor models in all samples and all waves

Sample		Model	χ^2 (df)	RMSEA [90%CI]	CFI	SRMR	AIC	Model Comparisons $\Delta\chi^2$ (Δdf)	*p* for $\Delta\chi^2$
ADSAT	Wave 1	A. One-factor	126.53 (9)	0.101 [0.086, 0.117]	0.92	0.048	19821		
		B. Two-factor [i]	102.51 (8)	0.096 [0.080, 0.113]	0.94	0.041	19799	A vs B = 24.02 (1)	<.001
		C. Two-factor [ii]	**50.14 (8)**	**0.064 [0.048, 0.081]**	**0.97**	**0.030**	**19747**	**A vs C = 76.39 (1)**	**<.001**
FTP-RBE	Wave 1	A. One-factor	123.12 (9)	0.150 [0.127, 0.174]	0.77	0.077	9632		
		B. Two-factor	82.04 (8)	0.128 [0.104, 0.154]	0.85	0.061	9592	A vs B = 41.08 (1)	<.001
		C. Two-factor	**62.59 (8)**	**0.110 [0.085, 0.136]**	**0.89**	**0.055**	**9573**	**A vs C = 60.53 (1)**	**<.001**
	Wave 2	A. One-factor	42.54 (9)	0.093 [0.066, 0.122]	0.92	0.049	7451		
		B. Two-factor	29.49 (8)	0.079 [0.050, 0.110]	0.95	0.037	7440	A vs B = 13.05 (1)	<.001
		C. Two-factor	**24.71 (8)**	**0.069 [0.039, 0.101]**	**0.96**	**0.041**	**7435**	**A vs C = 17.83 (1)**	**<.001**
	Wave 3	A. One-factor	34.05 (9)	0.095 [0.062, 0.130]	0.89	0.055	5327		
		B. Two-factor [iii]	29.56 (8)	0.093 [0.059, 0.131]	0.90	0.058	5325	A vs B = 4.49 (1)	.034
		C. Two-factor [iii]	**14.92 (8)**	**0.053 [0.000, 0.094]**	**0.97**	**0.040**	**5310**	**A vs C = 19.13 (1)**	**<.001**
Project KIDS	Wave 1	A. One-factor	74.74 (9)	0.129 [0.103, 0.157]	0.84	0.065	7615		
		B. Two-factor	56.31 (8)	0.117 [0.090, 0.147]	0.89	0.058	7598	A vs B = 18.43 (1)	<.001
		C. Two-factor	**35.35 (8)**	**0.088 [0.060, 0.119]**	**0.94**	**0.043**	**7577**	**A vs C = 39.39 (1)**	**<.001**
TEDS	Wave 3	A. One-factor	743.79 (9)	0.117 [0.110, 0.124]	0.87	0.058	101084		
		B. Two-factor	724.80(8)	0.122 [0.115, 0.130]	0.88	0.056	100657	A vs B = 18.99 (1)	<.001
		C. Two-factor	**313.88**	**0.080 [0.072, 0.087]**	**0.95**	**0.038**	**100656**	**A vs C = 429.92 (1)**	**<.001**
	Wave 4	A. One-factor	946.69 (9)	0.114 [0.108, 0.120]	0.89	0.056	133164		
		B. Two-factor	919.65 (8)	0.119 [0.113, 0.126]	0.90	0.054	133139	A vs B = 27.04 (1)	<.001
		C. Two-factor	**374.34 (8)**	**0.076 [0.069, 0.082]**	**0.96**	**0.034**	**132594**	**A vs C = 572.35 (1)**	**<.001**
WRRMP	Wave 1	A. One-factor	72.57 (9)	0.110 [0.088, 0.135]	0.91	0.056	9161		
		B. Two-factor	**39.12 (8)**	**0.082 [0.057, 0.108]**	**0.96**	**0.035**	**9129**	**A vs B = 33.44 (1)**	**<.001**
		C. Two-factor	49.47 (8)	0.095 [0.070, 0.121]	0.94	0.046	9139	A vs C = 23.09 (1)	<.001
	Wave 2	A. One-factor	51.03 (9)	0.096 [0.071, 0.122]	0.94	0.049	7953		
		B. Two-factor	40.74 (8)	0.089 [0.063, 0.118]	0.95	0.039	7944	A vs B = 10.29 (1)	<.001
		C. Two-factor	**29.54 (8)**	**0.073 [0.046, 0.101]**	**0.97**	**0.034**	**7933**	**A vs C = 21.49 (1)**	**<.001**
	Wave 3	A. One-factor	113.30 (9)	0.153 [0.129, 0.179]	0.86	0.067	7718		

Sample		Model	χ^2 (df)	RMSEA [90%CI]	CFI	SRMR	AIC	Model Comparisons $\Delta\chi^2$ (Δdf)	*p* for $\Delta\chi^2$
		B. Two-factor	103.94 (8)	0.156 [0.130, 0.183]	0.87	0.065	7711	A vs B = 9.36 (1)	.002
		C. Two-factor	**75.45 (8)**	**0.131 [0.105, 0.158]**	**0.91**	**0.052**	**7683**	**A vs C = 37.85 (1)**	**<.001**
	Wave 4	A. One-factor	39.69 (9)	0.098 [0.068, 0.131]	0.92	0.053	5630		
		B. Two-factor	**21.72 (8)**	**0.070 [0.035, 0.106]**	**0.96**	**0.040**	**5614**	**A vs B = 17.97 (1)**	**<.001**
		C. Two-factor	31.63 (8)	0.092 [0.060, 0.126]	0.94	0.047	5624	A vs C = 8.06 (1)	.004
	Wave 5	A. One-factor	52.62 (9)	0.115 [0.086, 0.146]	0.91	0.062	5754		
		B. Two-factor	**21.43 (8)**	**0.068 [0.034, 0.103]**	**0.97**	**0.047**	**5725**	**A vs B = 31.19 (1)**	**<.001**
		C. Two-factor	50.64 (8)	0.121 [0.090, 0.153]	0.91	0.059	5754	A vs C = 1.98 (1)	.159
	Wave 6	A. One-factor	53.66 (9)	0.116 [0.087, 0.147]	0.92	0.070	5664		
		B. Two-factor	**14.56 (8)**	**0.047 [0.000, 0.085]**	**0.99**	**0.047**	**5627**	**A vs B = 39.10 (1)**	**<.001**
		C. Two-factor	48.79 (8)	0.118 [0.087, 0.150]	0.93	0.066	5661	A vs C = 4.87 (1)	.027
	Wave 7	A. One-factor	16.68 (9)	0.059 [0.000, 0.102]	0.97	0.047	4036		
		B. Two-factor	**7.51 (8)**	**0.000 [0.000, 0.072]**	**1.00**	**0.037**	**4028**	**A vs B = 9.17 (1)**	**.002**
		C. Two-factor	16.43 (8)	0.065 [0.016, 0.111]	0.96	0.047	4037	A vs C = 0.26 (1)	.614

Note. [i] **Model B. Two-factor** tests the model proposed by Johnson et al. (2008). [ii] **Model C. Two-factor** tests the model suggested by the exploratory factor analysis of the ADSAT data. [iii] These models returned negative variances (i.e. Heywood cases) for the HomeZoo, BedRoutine, and HomeCalm items.

ings for item 5 *There is usually a television turned on somewhere in our home* (reversed) were notably different across samples, ranging from 0.07 (FTP-RBE) and 0.09 (Project KIDS) to 0.47 (TEDS), 0.37 (WRRMP) and 0.35 (ADSAT). We additionally examined modification indices for factor-item loadings by group. This post-hoc process showed that the three modifications with the largest expected change in χ^2 were those that freed the loading of item 5 on the Factor 1 (see Supplementary Table S14). Consequently, we released the constraint on the loading for this item, and tested a partial metric invariance model with the remaining five item loadings constrained to equivalence.

The partial metric invariance model (2B. in Table 3) fit the data significantly better than the full metric invariance model (2B vs 2A: $\Delta\chi^2$ (Δdf) = 68.36 (4)), and was not a worse fit to the data than the configural invariance model (2B vs 1: $\Delta\chi^2$ (Δdf) = 30.74 (12), *p*=.002; ΔRMESA = 0.002; ΔCFI = 0.002). We thus retained the partial metric invariance model and next tested scalar invariance by constraining all item intercepts to equivalence across the five samples, while retaining the partial metric invariance constraint. Fit statistics for scalar invariance (model 3, Table 3) show that this model was a worse fit to the data on all criteria compared with the partial metric invariance model ($\Delta\chi^2$ (Δdf) = 3320.65 (16), *p*<.001; ΔRMESA = 0.092; ΔCFI = 0.278). Since we allowed the loading for item 5 to vary, we next tested the possibility that the intercept for item 5 should also vary in a partial scalar invariance model (Model 3B). None of the comparative fit indices suggested this model was a better fit than the partial metric invariance model (Model 2B). Post-hoc examination of modification indices also showed no consistent pattern for possible item intercepts that could be freed. We therefore retained the partial metric invariance model.

Table 4 shows factor loadings and intercepts for each dataset for the retained partial metric invariance model for five samples. There is considerable variation in intercepts for some items across the five groups, after holding loadings constant for all but one item. For example, the intercepts for item 2, *You can't hear yourself think in our home* (reverse coded), range from 1.95 in the Project KIDS sample to 3.27 in the TEDS sample; similarly, for item 3. *It's a real zoo in our home,* intercepts range from 1.74 in the Project KIDS sample to 2.66 in the TEDS sample (N.B. because these item intercepts are allowed to vary by group, the model essentially reproduces item means reported in Tables S1-S10). The loadings for most other items have a smaller range, for example, 1.81-1.94 for item 4. *We are usually able to stay on top of things*. These differences in factor loadings indicate that response patterns vary across samples, potentially for reasons which are unrelated to differences in the latent construct under consideration (i.e. the confusion, hubbub and order of the home environment).

The R-square values reported in Table 4 provide additional information about the extent to which the variance in each item is captured by the final model. Of note is the low R^2 for two items. First, for item 5. *There is usually a television turned on somewhere in our home*, variance explained ranged from 0.2% (FTP-RBE), to 7% (ADSAT), to 14% (TEDS). Similarly, for item 1. *The children have a regular bedtime routine,* R^2 values were persistently low, with 3-5% of the variance explained by the factor model in all datasets. It is worth noting that both these items were first introduced when the short-form CHAOS was created, and did not appear in the original 15-item scale. For remaining items R^2 ranged from 16 to 77%.

Longitudinal invariance. We next tested longitudinal invariance of the scale for the three samples with repeated measures data. We used a similar approach to assess measurement invariance as that for the between-group models. We retained the same two-factor solution for the longitudinal models as for the between group models, and tested configural, metric, scalar, and residual invariance in repeated waves of data. We did not begin with the assumption that the partial metric invariance model retained in the between-group models would necessarily be warranted since differences between groups may not necessarily be observed within the same group over time. For within group models, we used the two available waves of data for the TEDS sample. While the FTP-RBE sample had three waves, we were able to include only wave 1 and wave 2 data as we encountered problems with Heywood cases in the wave 3 sample in the CFA step for both two-factor models. We used all seven waves for the WRRMP sample. Table 5 shows fit statistics for each step of longitudinal invariance testing for the three samples.

Interestingly, model fit comparisons for each step of invariance testing indicated that the TEDS sample was longitudinally invariant up to and including an equality constraint on the residuals (i.e. residual invariance). The FTP-RBE data showed borderline acceptable fit for the configural invariance model and no significant decrement in fit when factor loadings were constrained to equality for metric invariance ($\Delta\chi^2$ (Δ*df*) = 3.08 (4), *p*=.544). Fit was significantly worse, however, when the equality constraint on the intercepts was introduced (scalar invariance; $\Delta\chi^2$ (Δ*df*) = 32.09 (4), *p*<.001). On inspection of the intercepts across two waves, we noted the largest difference in item intercepts was for item 1 *the children have a regular bedtime routine* (wave 1 = 1.61; wave 2 = 1.95). We therefore tested a partial scalar invariance model allowing the intercept of this item to vary (model 3B. Table 5). This model was not a worse fit to the data compared with the metric invariance model ($\Delta\chi^2$ (Δ*df*) = 5.09 (3), *p*=.165), and a better fit than the residual invariance model. We therefore concluded partial scalar longitudinal invariance for the FTP-RBE data. Factor loadings, item intercepts and R-squared for each item for the FTP-RBE samples are included in Table S15 in the supplementary material.

In a contrast to these results, the fit statistics for the WRRMP sample were very borderline for the configural invariance step, no worse for the metric invariance step ($\Delta\chi^2$ (Δ*df*) = 35.21 (24), *p*=.065) – and a slight improvement if judged by ΔRMESA = 0.015 – though progressively worse for both scalar and residual invariance. We report factor loadings and item intercepts for the metric invariance step for the WRRMP data in Table S16 in the supplementary

Table 3. Model fit statistics for measurement invariance tests including one wave from each of five samples

Model	χ^2 (df)	RMSEA [90% CI]	CFI	SRMR	AIC	Model Comparisons $\Delta\chi^2$ (Δdf)	*p* for $\Delta\chi^2$
1. Configural	551.96 (40)	0.077 [0.071, 0.083]	0.957	0.035	177424		
2A. Metric	651.06 (56)	0.070 [0.065, 0.075]	0.950	0.041	177492	1 vs 2A = 99.10 (16)	<.001
2B. Partial Metric	**582.70 (52)**	**0.069 [0.064, 0.074]**	**0.955**	**0.037**	**177431**	**1 vs 2B = 30.74 (12)**	**.002**
3A. Scalar [i]	1343.26 (68)	0.093 [0.089, 0.097]	0.892	0.059	178160	2B vs 3 = 760.57 (16)	<.001
3B. Partial Scalar [ii]	1074.36 (64)	0.085 [0.081, 0.090]	0.915	0.051	177899	3A vs 3B = 268.90 (4) 2B vs 3B = 491.66 (12)	<.001 <.001

Note. [i] The scalar invariance model allowed for partial metric invariance – i.e. factor loadings of the TV item were allowed to vary across groups. [ii] The partial scalar invariance model allowed the intercept of the TV item to vary. Retained model is in bold.

Table 4. Factor loadings, intercepts and R-square values for each item and group for the retained partial metric invariance model

	ADSAT			Florida Twin Study			Project Kids			WRRMP			TEDS		
Item	Loading	Intercept	R^2	Loading	Intercept	R^2	Loading	Intercept	R^2	Loading	Intercept	R^2	Loading	Intercept	R^2
Factor 1															
2. HomeNoise	0.86	2.63	.62	0.86	2.25	.46	0.86	1.95	.51	0.86	2.51	.70	0.86	3.27	.68
3. HomeZoo	0.92	2.13	.70	0.92	1.83	.67	0.92	1.74	.76	0.92	2.32	.73	0.92	2.66	.65
5. HomeTV [i]	0.35	2.72	.07	0.07	3.63	.002	0.09	3.71	.006	0.37	3.09	.09	0.47	3.26	.14
Factor 2															
1. BedRoutine	0.15	1.43	.04	0.15	1.61	.05	0.15	1.84	.03	0.15	1.46	.05	0.15	1.38	.05
4. HomeControl	0.32	1.81	.22	0.32	1.81	.19	0.32	1.94	.16	0.32	1.92	.24	0.32	1.89	.18
6. HomeCalm	0.72	2.55	.60	0.72	2.25	.76	0.72	2.11	.77	0.72	2.65	.65	0.72	2.85	.60

Note. Standardized latent factors (M=0; SD=1). [i] Loadings are allowed to vary for this item, all other items loadings are constrained to equality.

Table 5. Model fit statistics for longitudinal measurement invariance tests for TEDS, FTP-RBE and WRRMP data

Model	χ^2 (df)	RMSEA [90% CI]	CFI	AIC	SRMR	Model Comparisons $\Delta\chi^2$ (Δdf)	*p* for $\Delta\chi^2$
TEDS [i]							
1. Configural Invariance	688.22 (16)	0.077 [0.073, 0.082]	0.954	233251	0.036		
2. Metric Invariance	701.32 (20)	0.070 [0.065, 0.074]	0.953	233255	0.037	1 vs 2 = 13.10 (4)	.011
3. Scalar Invariance	716.69 (24)	0.064 [0.060, 0.068]	0.953	233263	0.037	2 vs 3 = 15.37 (4)	.004
4. Residual Invariance	**738.13 (30)**	**0.058 [0.054, 0.062]**	**0.952**	**233272**	**0.038**	**3 vs 4 = 21.44 (6)**	**.002**
FTP-RBE [ii]							
1. Configural Invariance	87.30 (16)	0.094 [0.076, 0.114]	0.922	17008	0.049		
2. Metric Invariance	90.38 (20)	0.084 [0.067, 0.102]	0.923	17003	0.051	1 vs 2 = 3.08 (4)	.544
3. Scalar Invariance	122.47 (24)	0.091 [0.075, 0.107]	0.892	17027	0.061	2 vs 3 = 32.09 (4)	<.001
3B. Partial Scalar	**95.48 (23)**	**0.079 [0.063, 0.096]**	**0.920**	**17002**	**0.052**	**2 vs 3B = 5.09 (3)**	**.165**
4. Residual Invariance	115.70 (29)	0.077 [0.063, 0.092]	0.905	17010	0.064	3B vs 4 = 20.22 (6)	.003
WRRMP [iii]							
1. Configural Invariance	301.95 (56)	0.103 [0.091, 0.114]	0.935	45832	0.049		
2. Metric Invariance	**337.16 (80)**	**0.088 [0.078, 0.098]**	**0.933**	**45819**	**0.055**	**1 vs 2 = 35.21 (24)**	**.065**
3. Scalar Invariance	548.91 (104)	0.101 [0.093, 0.110]	0.883	45983	0.076	2 vs 3 = 211.75 (24)	<.001
4. Residual Invariance	792.71 (140)	0.106 [0.099, 0.113]	0.829	46155	0.097	3 vs 4 = 243.81 (36)	<.001

Note.[i] TEDS comprised two waves of data collection. [ii] The FTP-R data comprised two waves of data. Wave 3 was excluded because of problems with negative observed variable variances (i.e. Heywood cases) in confirmatory factor analyses. [iii] The WRRMP comprised seven waves of data.

material. Since there were no clearly apparent patterns in the differences in intercepts across the seven waves of data collection, it is difficult to identify what may have caused model misfit in the scalar invariance step. It is possible that slight variations in intercepts of multiple items, or at different waves, could contribute to the decrement in model fit from metric to scalar invariance.

Reliability statistics. Given the six items in the scale are most often used as a sum or average score, we computed two internal reliability statistics for descriptive purposes, and to collate this information in one location for future reference: Cronbach's alpha (α) and omega (hierarchical, ω_h; see Table 1). While Cronbach's alpha is reported in almost all published papers using CHAOS data, whether the assumptions for alpha are met is not often identified. These assumptions include: unidimensionality of the scale, tau-equivalence (i.e. equivalence of factor loadings), and uncorrelated item residuals (Kenny, 1979; Revelle, n.d.). First, the results of the EFA and CFAs suggested that the assumption of unidimensionality for the six-item scale was violated in each dataset. A two-factor solution was indicated by the EFA, and CFA results showed a significant improvement in model fit for a two-factor compared with a one-factor model each case. Second, the factor loadings reported in Table 4 suggest that the scale is not tau-equivalent, with loadings varying considerably across items. We tested a model that forced item loadings to be equivalent and found that model fit was significantly worse than when item loadings were allowed to vary (i.e. compared with Model A. Table 2). Using the ADSAT data as an example, the difference in fit showed the equal factor loadings model to be a worse fit to the data than the model that allowed loadings to vary ($\Delta\chi^2$ (Δdf) = 641.55 (5); ΔRMESA = 0.103; ΔCFI = 0.436). We found the same decrease in fit across all the datasets when testing this assumption. Fit statistics for these models are in Supplementary Table S17.

Finally, the one-factor CFAs reported in Table 2 were estimated with uncorrelated residuals as per the default options in the *lavaan* package (Rosseel, 2012). Since these models showed poor fit to the data in each case, we used information gleaned from modification indices to identify whether correlations amongst observed item residuals would improve model fit – i.e. were the residuals of the items truly uncorrelated? Modification indices greater than χ^2 =10 for the one-factor CFA models are in Supplementary Table S18. The chi-square values represent the estimated size of the improvement in model fit should the modification be made to the CFA. Each modification has one degree of freedom. For each dataset, the modification indices flagged between two and 10 item residual correlations that could be added to the one-factor models that would significantly improve model fit, indicating that the assumption of uncorrelated residuals was violated.

Since the assumptions of alpha were violated, we also computed omega (hierarchical, ω_h) as a measure of internal consistency reliability. Omega is arguably a more appropriate indicator of reliability because it allows for different factor loadings of items (McNeish, 2018), although the assumption of unidimensionality will still be violated if a one-factor model is inappropriate. Estimates of omega are also reported in Table 1 and range from ω_h = 0.29-0.63, again indicating that internal consistency for the six-item scale was poor across samples.

Predictive validity. Given the inconclusive nature of the measurement invariance tests, and the poor internal consistency reliability of the scale, it is difficult to examine the predictive validity of the short-form CHAOS. Nonetheless, if only to provide a demonstration of how correlation coefficients can vary when different methods of generating a composite score are used, and when a scale has low reliability. We opted to estimate correlations between two different configurations of the six items and available criterion variables for each dataset. The criteria we selected were a socioeconomic status proxy and any available academic achievement variables. We selected the same wave of data as that selected for between-group measurement invariance tests described above. Since we conducted multiple tests, we used a conservative significance threshold of p <.01 to assess which correlations were statistically significantly different from zero.

First, we generated a single-variable factor score using all six items, following the most common use of the scale. Secondly, we generated factor scores for two variables based on the two-factor solution with the best-fitting model. Specifically, these factors comprised three variables each and were termed *disorder* (items 1, 4 and 6), and *noise* (items 2, 3, and 5). Using the *psych* package in R (Revelle, 2022), factor scores were generated separately for each configuration of items, producing variables with M=0 and SD=1. Table 6 shows correlations between factor scores and criterion variables for one- and two-factor configurations of items. For comparative purposes, we generated the same correlations between the latent factor CHAOS and the criteria by extending the CFA models to include the criteria (reported in Table 2). We note that the size and direction of correlations were no different using this approach (see Supplementary Table S19).

Correlations with achievement were either negative (as expected), or negligible, varying by sample and whether a one-factor or two-factor combination of items was used. For four of the five datasets – the ADSAT, FTP-RBE, Project KIDS and TEDS – the correlations reported in Table 6 supported our prediction that higher levels of parent-reported confusion, hubbub and disorder in homes would be negatively associated with measures of academic achievement. In these four datasets, correlations between a one-factor CHAOS measure and achievement ranged between r = -.11 to -.21 (p <.001). Interestingly, when the six variables were separated into two factors, only the *Noise* factor consistently correlated with achievement, while the *Disorder* factor did not. The most notable exception was the Project KIDS data where the *Disorder* factor correlated negatively with both English and Math grades (r = -.20 / -.17 respectively, p <.001), even though the correlation between Math grades and the one-factor CHAOS was small and not significantly different from zero (r = -.06). By contrast, the correlations between one-factor CHAOS variable and the two factors (*Noise* and *Disorder*) and several criteria were small

Table 6. Correlations between one-factor and two-factor CHAOS, socioeconomic status (SES) and academic achievement criterion variables.

Study	Correlated variable	One-factor	Two-factor		
		Chaos	Noise	Disorder	Factor correlation
ADSAT	SES	-.12***	-.11***	-.09***	.45***
	Grade 3 Reading	-.11***	-.12***	-.04	
	Grade 3 Math	-.07	-.07	-.03	
FTP-RBE Wave 1	SES	-.10	-.15***	-.09	.21***
	FCAT Reading 2011-12	-.16	-.26**	-.06	
	FCAT Reading 2012-13	-.21***	-.27***	-.08	
Project KIDS	SES	-.09	-.23***	-.07	.29***
	English Language Arts Grade	- .18***	-.20***	-.20***	
	Math Grade	- .06	-.07	- .17***	
WRRMP Wave 2[i]	SES [ii]	-.04	-.03	.01	.20***
	PIAT passage comprehension	.01	.00	.01	
	WRMT passage comprehension	-.04	-.05	-.00	
	WJ Calculation	-.06	-.06	-.03	
	WJ applied problems	-.01	-.02	-.04	
	WJ quantitative concepts	-.03	-.04	.00	
TEDS Wave 4	SES	-.23***	-.24***	-.05***	.33***
	English National Curriculum Assessment [iii]	-.16***	-.16***	-.07	
	Math National Curriculum Assessment [iii]	-.11***	-.12***	-.04	

Note. Correlations are between factors and criterion variables. $^{**}p$ <.01; $^{***}p$ <.001. [i] For the WRRMP Wave 2 data we use the same configuration of items as the remaining datasets for the two factor models, notwithstanding the better fit of the alternative model in wave 1 data. Interestingly correlations remained non-significant with the alternative item configuration reported in Johnson et al. (2008). [ii] Academic assessments at age 7.

and non-significant for the WRRMP data. This dataset contained several high-quality measures of both reading comprehension and mathematics sub-domain skills (calculation, applied problems, and quantitative concepts) yet none correlated with the CHAOS.

Finally, and in accordance with the usual use of the CHAOS items, we generated composite variables and correlated these with the criteria. We tested both combinations of the six items, i.e. a single variable averaging across the six CHAOS items, and two variables using averages of the same three items as used in the factor score models. Table 7 shows correlations between average CHAOS scores and criterion variables. Patterns of results were similar when composite variables were created by averaging across items (see Table 7). One exception was noted in the WRRMP data, where the WRMT passage comprehension measure correlated negatively with the composite score of six items (r = -.16) and with the *Noise* composite (r = -.20), but not the *Disorder* composite.

Correlations between SES and CHAOS also differed by sample, whether one or two composites were used, and method of creating variables (factors or averages). When average scores were used (Table 7), SES correlated negatively with the one-variable CHAOS across all datasets (r= -.15 to -.32), though the WRRMP correlation was significant only at the p <.01 level. Similar to the academic achievement variables, when CHAOS was separated into the two composites, SES correlated more consistently with the *Noise* composite rather than the *Order* composite. When factor scores were used to generate the CHAOS variables (Table 6), patterns of correlations with SES were similar, but notably smaller in all datasets, and correlations between the one-factor CHAOS and SES were significantly different from zero in only the ADSAT (r= -.12) and the TEDS data (r= -.23). In the Project KIDS and FTP-RBE samples, correlations with *Noise* factor and CHAOS were significantly different from zero (r= -.15 / -.23; p<.001) even though the correlations with the six-item factor were not.

Discussion

The central aim of this study was to examine the measurement properties of the short form of the Confusion, Hubbub and Order Scale (CHAOS) with a goal to provide some recommendations on the use of the scale, both in pre-existing datasets and new research. On the whole, our results indicate that the six items in the short-form CHAOS are not reliable and valid *enough* to capture variability in the quality of home environments across different contexts, age ranges of child study participants, and time. Ideally, reduction of a long-form to a short-form scale should be accompanied by evidence that the short form itself is a) reliable, b) valid, and c) captures the breadth of the construct indicated by a long-form (Clark & Watson, 1995, 2019; Smith et al., 2000). Because these steps were not documented for the short-form CHAOS, this study provides some of the information necessary to guide the use of the scale in future applied research. The motivation for this

Table 7. Correlations between CHAOS average composites, socioeconomic status (SES) and academic achievement criterion variables

Study	Correlated variable	One variable	Two variables		
		Chaos	Noise	Disorder	Composite score correlation
ADSAT	SES	-.21***	-.23***	-.10***	.41***
	Grade 3 Reading	-.16***	-.20***	-.04	
	Grade 3 Math	-.12***	-.15***	-.04	
FTP-RBE Wave 1	SES	-.21***	-.23***	-.10	.27***
	FCAT Reading 2011-12	-.24***	-.30***	-.06	
	FCAT Reading 2012-13	-.25***	-.29***	-.10	
Project KIDS	SES	-.21***	-.25***	-.07	.31***
	English Language Arts Grade	-.26***	-.22***	-.20***	
	Math Grade	-.15***	-.07	-.18***	
WRRMP Wave 2	SES	-.15**	-.19***	-.03	.39***
	PIAT passage comprehension	-.09	-.12	-.01	
	WRMT passage comprehension	-.16***	-.20***	-.04	
	WJ Calculation	-.10	-.13	-.03	
	WJ applied problems	-.10	.14	.00	
	WJ quantitative concepts	-.07	-.12	.02	
TEDS Wave 4 [i]	SES	-.32***	-.38***	-.09***	.36***
	English National Curriculum assessment	-.21***	-.23***	-.09***	
	Math National Curriculum Assessment	-.15***	-.17***	-.06***	

Note. Composite variables created by averaging six items for one-factor CHAOS; three items each for *Noise* and *Disorder*. ***p* <.01; ****p* <.001. [i] English and math assessments at age 7.

study was additionally underpinned by the growing calls for more rigorous approaches to the development and evaluation of survey measures, and iterative reconsiderations of conceptual clarity as necessary precursors for advancing educational and psychological sciences (Bringmann et al., 2022; Flake, 2021; Smaldino, 2019; van Dijk et al., 2021).

To date internal consistency reliability estimates (i.e. coefficient alpha) have been the only consistently documented evidence of scale reliability for the short-form CHAOS. Estimates of alpha are universally low (i.e. <.70 in all cases) in the samples included in this study, and in other samples reported in the published literature, suggesting that the internal consistency of the items may not be sufficient for combining them in a single composite (McNeish & Wolf, 2020). In addition, the results of our assumption testing for calculating alpha demonstrated that all the assumptions (unidimensionality, tau-equivalence and uncorrelated residuals) were violated in each of the datasets used in this study. These results suggest the calculated alphas cannot be interpreted with confidence, particularly when the scale is not unidimensional.

Creating a single composite from several items assumes that a scale is unidimensional (i.e. captures a single latent domain; McNeish & Wolf, 2020). In all but one of the published studies using the short-form CHAOS, the six items were either summed or averaged to compute a single composite 'chaos' variable. Previous published evidence for the short-form CHAOS (Johnson et al., 2008) and our own exploratory factor analysis suggested that the six items better (though not perfectly) represent two latent domains. An interpretational difficulty arose however, with the identification of two different patterns of item-to-factor loadings in each of these EFA samples: a possibility that was not originally considered in our preregistration of the study. These different item-to-factor patterns, along with poor model fit in confirmatory factor models in several samples (see Table 3), and poor internal consistency reliability estimates, support our conclusion that the short-form CHAOS may not satisfactorily capture a single dimension of home environment quality suggested by theoretical descriptions (Bradley, 2015; Bronfenbrenner & Ceci, 1994; Bronfenbrenner & Evans, 2000; Matheny et al., 1995).

Nonetheless, and notwithstanding these conceptual problems, our hypothesis that a two-factor dimensional structure would fit the CHAOS data better than a single dimension was broadly supported in all samples. In four of the five datasets, the two-factor structure suggested by an EFA using a subsample of participants in the ADSAT data was the best fit. By contrast, the best-fitting model in the WRRMP data varied by wave: in five of seven waves the alternative two-factor structure identified in previous analyses of these data (Johnson et al., 2008) was a better fit. Despite these findings, however, the fit of the confirmatory factor models in some samples and waves was still poor according to all model fit criteria. In particular, none of the tested models returned adequate model fit statistics in wave 3 of the WRRMP data, nor wave 1 of the FTP-RBE data. We also encountered negative variances in wave 3 of the FTP-RBE (i.e. Heywood cases). These results suggest that participants completing the short-form CHAOS items may respond inconsistently by group or data collection wave, rendering the meaning of a composite of the six items unclear.

Furthermore, results of between-group measurement invariance analyses indicated that the short-form CHAOS was non-invariant across the five included samples, thus supporting our second hypothesis. Specifically, while the configural invariance model incorporating five samples was an adequate fit to the data (see Table 3, model 1.), neither the full metric invariance, nor the full scalar invariance models were acceptable according to multiple model fit criteria (Rutkowski & Svetina, 2014; West et al., 2012). The retained model was a partial metric invariance model, which constrained the factor loadings of five items to equality, and allowed the loading of one item to vary (item 5. *There is usually a television turned on somewhere in our home*). While this model was acceptable, scalar invariance was not supported, indicating that the intercepts of the items differed too greatly across samples for them to be constrained to equality. Since we allowed the loading of item 5 to vary in the partial metric invariance model, we additionally tested a partial scalar invariance model, allowing the intercept of item 5 to vary. This partial scalar invariance model was also unacceptable according to (change in) fit criteria. Because we included five samples in the measurement invariance analyses, it is neither easy nor straightforward to identify which samples and/or items may have been driving the scalar non-invariance – it could be a combination of any number of items across some or all samples.

Interestingly, our hypothesis of longitudinal non-invariance was not supported in all studies with multi-wave data collection. In particular, measurement invariance analyses showed that the two-factor model was invariant over the two waves of data in the TEDS sample, up to and including invariance of the residuals. Secondly, allowing for variation in the intercept of the bedtime item, two waves of data in the FTP-RBE were longitudinally invariant. This finding, however, is tempered by our inability to include the third wave of FTP-RBE data due to estimation problems in the CFA models. By contrast, we could not conclude that the CHAOS measure in the WRRMP sample was invariant. The best fitting model in these data was the metric invariance model (equal loadings), though scalar invariance (equal intercepts) was not supported. Again, the multiple waves of data in the WRRMP make it difficult to pinpoint a possible reason for this finding. Nonetheless, closer examination of the items included in the short-form CHAOS could shed some light on possible explanations and candidates for future revisions of the scale.

It is possible that the non-invariance between and within samples was driven by the two items that were added to the short-form version of CHAOS but did not appear in the original validated long form. These two items, 5. *There is usually a television turned on somewhere in our home* and 1. *The children have a regular bedtime routine,* were the worst performing items across all samples, with low factor loadings and <10% of the variance of the items explained by any model (with the exception of item 5 in the TEDS data = 14% of variance explained). In terms of face validity, item 5 has become particularly dated in western cultures in the 25 years since the short-form CHAOS was proposed. For example, the data in the TEDS project was collected in the late 1990s, whereas the most recent data collection, Project KIDS, was in 2017. Compared with the 1990s, 21st century middle-class families (and children) now have access to an abundance of portable electronic devices, including smartphones, tablets, and laptops. Children and adults have access to headphones, volume control, voice-activated commands and individualized options. If the television item intended to capture ambient noise within a household, it may be outdated. If the item intended to capture parents' lack of control over children's media consumption, again, the item will likely no longer capture the range of digital media currently available to children and adolescents (Graafland, 2018).

Secondly, while the face validity of the bedtime routine item might be acceptable for samples of very young children, this routine might not be applicable to older children and adolescents. Our finding of non-invariant intercepts for this item in the FTP-RBE sample supports this idea: in wave 1, children in this study were ~11 years old, and in wave 2 they were ~13 years old. The increasing item intercept indicates the parents were less likely to agree that children had a regular bedtime routine as they aged. By comparison we did not encounter this same problem in the TEDS sample where the ages of the children were 3 years and 4 years for waves 3 and 4 respectively. The bedtime routine question therefore may capture different expectations for middle childhood compared with toddlerhood, developmental changes in sleep patterns, or child or parent personality, rather than an aspect of household management and order. Unpacking the assumptions embedded in the question as it relates to variability in household order and routine raises additional questions: Is it problematic or damaging if older children lack a strictly adhered-to bedtime routine every evening? Is the amount and nature of sleep itself a better predictor of positive childhood development than regularity in bedtimes (e.g. Dewald et al., 2010)? If the CHAOS scale is to be applied in research spanning early childhood to mid-adolescence, as is currently the case, these questions, and the face validity of all the items, should be examined in the light of advancements to developmental theory, and changes to family life that have occurred since the mid-1980s when the scale was first developed in a sample of infants and toddlers.

Thirdly, the contrasting results of the two exploratory factor analyses (i.e. our own and that reported by Johnson et al., 2008) suggest that item 6 *the atmosphere in our house is calm* potentially lacks conceptual clarity (Borsboom et al., 2004; Bringmann et al., 2022). In the WRRMP data, this item loaded with others representing 'quietness of the household' (Johnson et al., 2008), whereas in the ADSAT data this item cross-loaded highly onto both factors, suggesting that respondents may have varying interpretations of what it means to have a calm atmosphere in the home. Furthermore, whether or not a calm atmosphere is representative of poor household environments is arguable, and potentially tied to cultural norms, thus perhaps leading to the inconsistent properties of this item across datasets. While the issues described here in relation to items 1, 5 and 6 could explain the findings of measurement non-invariance, either between groups or within groups, it is not possible to arrive at definite answers using the measurement invariance approach described in this study. The inconclusive results of the WRRMP longitudinal models are a case in point here: while we can conclude that scalar invariance is not supported, we cannot say exactly why. Future researchers would therefore need to examine face validity, relevance, conceptual clarity and theoretical and cultural appropriateness of items before additional scale development and evaluation work can be undertaken.

Finally, our investigation of the predictive validity of the short-form CHAOS scale was limited by the finding of measurement non-invariance. However, using two different approaches to collating variables (i.e. factor scores or averaged composites) and comparing correlations with socioeconomic status and academic achievement variables is nonetheless instructive, particularly given the usual use of the scale. Using either approach to combining items, higher ratings of CHAOS correlated with poorer academic achievement in four of five samples. Factor-score correlations were generally smaller than those observed when items were averaged to create composite variables. In the WRRMP sample, CHAOS did not consistently correlate with any of the five reading or math assessments.

Similarly, while SES and CHAOS were negatively correlated in general across the five samples, of note are the differences in the strength of correlations (and their statistical significance) when the average score was used rather than the factor score. In all samples, correlations using an average CHAOS composite were larger than those using the factor score. For the WRRMP and Project KIDS samples, correlations with the factor score were not significantly different from zero, whereas the correlations with the average composite were significantly different from zero. It is worth reiterating that measurement error of observed items is retained in composite variables, and can subsequently inflate or reduce covariations in unpredictable ways (Cole & Preacher, 2014; McNeish & Wolf, 2020) – as we have observed in these comparisons. While this problem can be somewhat rectified by the use of factor scores which allow differential weighting of items comprising each factor, if the observed items do not reliably capture the underlying theoretical construct, a factor score approach does not completely resolve the measurement problems (Hancock, 2003; Rhemtulla et al., 2020). The only resolution in this and many other cases is careful and considered development and renewal of items in the light of theoretical construct of interest.

Recommendations. Since there are multiple studies that have collected data on the short form CHAOS over the past 20 years, and several of these data sources are now accessible to researchers for secondary analyses, we provide some tentative recommendations on the use of the six CHAOS items. Given the finding that a two-factor solution demonstrated better fit to the data than the commonly used one-factor model in all cases, we recommend that future researchers should investigate whether it is more appropriate to use a two-factor configuration of the six observed items with factors termed *Noise* and *Disorder.* These two subfactors arguably reflect the theoretical conceptualisation of household ‘chaos’ described in the literature, and in the description of the original 15-item CHAOS (Matheny et al., 1995). In particular, the original scope of the theoretical domain included high levels of noise within a home, and disorganization, as two of the defining features of household ‘chaos’ (Evans, 2006; Matheny et al., 1995). While the two factors suggested by this study do not adequately cover other aspects of the original theoretical definition, such as the numbers of visitors coming and going, frenetic activity, and clutter within a home, none of the six items, when treated individually, capture these concepts.

Results using the two-factor approach may also be compared with the one-factor approach commonly reported in the literature, as we have done in this paper. To this end, and for comparative purposes, we fitted between-group measurement invariance models for a one-factor model using data from the same five samples reported in this paper. Supplementary Table S20 shows the fit statistics for this model. The results demonstrate the poor fit of the one-factor model in the configural invariance step (χ^2 (df) = 1322.10 (45); RMESA = 0.115; CFI = 0.89), and that neither full metric nor full scalar invariance was supported using this method of combining items.

Finally, we would also suggest that using factor scores, or structural equation models, allowing items in the subscale(s) to be differentially weighted, is more appropriate than using sum or average scores – particularly given the variation in correlation coefficients using each approach. These recommendations, however, should not be taken as rules, and should not preclude researchers from carefully examining the properties of the items in samples not included in this study.

Limitations

A major limitation of the analyses presented here is the non-definitive nature of the information obtained from measurement invariance tests when more than two samples are included. While we suggest that the item relating to television may be driving between-group metric non-invariance, other items could also be contributing to this result. A second limitation is the differing ages of the children included in each sample. While we made efforts to select samples with similarly aged children, this was not always possible due to the secondary data accessed for the study. Mean age ranged from 4 years in the TEDS sample, to 11 years in both the FTP-RBE and Project KIDS samples. Differing ages of children when parents respond to the items could drive differential response patterns across samples. Nonetheless, if this is the case, it is further evidence that the short-form CHAOS is not as broadly applicable across childhood and adolescence as it is intended to be.

Second the samples used in this study were all reasonably similar in terms of socioeconomic status, with the possible exception of the FTP-RBE sample. While each of the studies made every effort to recruit samples that were representative of a country, or region within a country, there is an overrepresentation of middle class, white and educated parents who participate in these (and other) survey-based studies of children. We therefore cannot easily extrapolate our findings to other contexts, countries, languages, or populations. The Parenting Across Cultures (PAC) study provides some evidence that not all the items in the short-form CHAOS may be relevant in cultures other than Western, Educated, Industrialized, Rich and Democratic (WEIRD) ones largely represented in the samples used in this study. In particular, the PAC study dropped the television item from analyses using the CHAOS due to poor face validity, appropriately recognising that this item was likely not a good indicator of household order and routine for the participants in that study drawn from middle-income countries.

Furthermore, four of the five samples used in this study come from twin studies. For testing predictive validity, we selected one twin from each pair. We could have used multilevel structural equation modeling to make use of all available achievement data from twins nested within families. However, for the purposes of examining whether or not the construct predicts outcomes in expected ways, we opted for simplicity so as to make the results as plain as possible to as wide a readership as possible. It is unlikely that a more complex analytic approach would alter our conclusions about predictive validity, and since CHAOS is measured only once per family all the available CHAOS data is used in each case.

Finally, because the analyses are largely data-driven, the analytic choices, and the order in which different steps were undertaken in this study were affected by researcher degrees of freedom (Gelman & Loken, 2013). It would be possible to attempt different analyses and obtain different results, for example, is a one-factor solution acceptable if the television item is omitted? Or both the television and bedtime items? These different choices, however, would not get us closer to the main object of interest, which is to identify whether the six items in the short-form CHAOS are valid and reliable measure of the quality of home environments. Future work may consider these and other options within a broader program of scale development and renewal.

Conclusion

Studies of the links between the nature of home environments and childhood development are decades old (e.g. Bronfenbrenner, 1981; Elardo et al., 1977). The original 15-item CHAOS measure clearly identified the aspects of home environments it was intended to capture. These included household disorder, high ambient noise, and lack of routine (Matheny et al., 1995) and items were developed from a wealth of earlier theorizing about how variation in home environments might relate to different aspects of early childhood development. However, the results we report here do not provide strong evidence that the short-form CHAOS adequately captures this broad and theoretically consistent construct. The rationale for selecting the six items is arguably clear: in terms of face validity and relevance, the items do cover the scope of the original construct, albeit in a more limited way. However, our findings indicate that the short form items should now be reconsidered and the scale revised in the light of more contemporary theory and contexts (e.g. Clark & Watson, 1995, 2019). Perhaps the best place to begin this process would be with a systematic re-evaluation and update of the 15 items in the original version of the Confusion, Hubbub and Order Scale.

Contributions

Conception and design – Sally A. Larsen and Callie W. Little

Acquisition of data – Kathryn Asbury, William L. Coventry, Sara A. Hart and Stephen A. Petrill

Analysis and interpretation of data – Sally A. Larsen and William L. Coventry

Drafted and revised the article – Sally A. Larsen

Final article approval – Sally A. Larsen, Kathryn Asbury, William L. Coventry, Sara A. Hart, Callie W. Little, and Stephen A. Petrill

Funding Information

Academic Development Study of Australian Twins

This research was supported by two Australian Research Council Discovery Project Grants: DP 120102414 (2012-2014) and DP 150102441 (2015-2018). Access to the sample was facilitated by Twins Research Australia, a national resource supported by a Centre of Research Excellence Grant (ID: 1079102), from the National Health and Medical Research Council.

Florida Twin Project on Reading, Behaviour and Environment

This work was supported by Eunice Kennedy Shriver National Institute of Child Health & Human Development Grant 5P50HD052120-08

Project KIDS

This work was supported by Eunice Kennedy Shriver National Institute of Child Health & Human Development Grants R21HD072286, P50HD052120, and R01HD095193.

Western Reserve Reading and Math Project

This work was supported by Eunice Kennedy Shriver National Institute of Child Health & Human Development Grants HD38075 and HD46167

Twins Early Development Study

TEDS is supported by a program grant from the UK Medical Research Council (MR/M021475/1 and previously G0901245), with additional support from the US National Institutes of Health (AG046938)

Competing Interests

None.

Data Accessibility Statement

ADSAT data is accessible by application to the first and third authors

Available at the LDBase repository https://ldbase.org/ are:

- FTB-RBE (doi:10.33009/ldbase.1624451991.3667)
- Project KIDS (doi: 10.33009/ldbase.1619716971.79ee) and
- WRRMP (doi: 10.33009/ldbase.1643647076.d4b2)

TEDS data available by agreement with the Kings' College London https://www.teds.ac.uk/researchers/teds-data-access-policy

Submitted: September 25, 2022 PDT, Accepted: May 02, 2023 PDT

References

Asbury, K., Dunn, J. F., Pike, A., & Plomin, R. (2003). Nonshared environmental influences on individual differences in early behavioral development: A monozygotic twin differences study. *Child Development*, *74*(3), 933–943. https://doi.org/10.1111/1467-8624.00577

Asbury, K., Wachs, T. D., & Plomin, R. (2005). Environmental moderators of genetic influence on verbal and nonverbal abilities in early childhood. *Intelligence*, *33*(6), 643–661. https://doi.org/10.1016/j.intell.2005.03.008

Australian Curriculum Assessment and Reporting Authority. (2017). *NAPLAN achievement in reading, writing, language conventions and numeracy: National report for 2017*. https://www.nap.edu.au/results-and-reports/national-reports

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295x.111.4.1061

Bradley, R. H. (2015). Constructing and adapting causal and formative measures of family settings: The home inventory as illustration. *Journal of Family Theory & Review*, *7*(4), 381–414. https://doi.org/10.1111/jftr.12108

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science*, *31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Bronfenbrenner, U. (1981). *The Ecology of Human Development: Experiements by Nature and Design*. Harvard University Press. https://doi.org/10.2307/j.ctv26071r6

Bronfenbrenner, U. (1986). Ecology of the family as a context for human development: Research perspectives. *Developmental Psychology*, *22*(6), 723–742. https://doi.org/10.1037/0012-1649.22.6.723

Bronfenbrenner, U., & Ceci, S. J. (1994). Nature-nuture reconceptualized in developmental perspective: A bioecological model. *Psychological Review*, *101*(4), 568–586. https://doi.org/10.1037/0033-295x.101.4.568

Bronfenbrenner, U., & Evans, G. W. (2000). Developmental science in the 21st century: Emerging questions, theoretical models, research designs and empirical findings. *Social Development*, *9*(1), 115–125. https://doi.org/10.1111/1467-9507.00114

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In W. Damon, R. M. Lerner, & R. M. Lerner (Eds.), *Handbook of Child Psychology*.

Butcher, L. M., & Plomin, R. (2008). The nature of nurture: A genomewide association scan for family chaos. *Behavior Genetics*, *38*(4), 361–371. https://doi.org/10.1007/s10519-008-9198-z

Byrne, B. M. (2012). *Structural Equation Modeling with Mplus: Basic Concepts, Applications and Programming*. Routledge.

Chang, L., Lu, H. J., Lansford, J. E., Bornstein, M. H., Steinberg, L., Chen, B.-B., Skinner, A. T., Dodge, K. A., Deater-Deckard, K., Bacchini, D., Pastorelli, C., Alampay, L. P., Tapanya, S., Sorbring, E., Oburu, P., Al-Hassan, S. M., Di Giunta, L., Malone, P. S., Uribe Tirado, L. M., & Yotanyamaneewong, S. (2019). External environment and internal state in relation to life-history behavioural profiles of adolescents in nine countries. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1917), 20192097. https://doi.org/10.1098/rspb.2019.2097

Chang, L., Lu, H. J., Lansford, J. E., Skinner, A. T., Bornstein, M. H., Steinberg, L., Dodge, K. A., Chen, B. B., Tian, Q., Bacchini, D., Deater-Deckard, K., Pastorelli, C., Alampay, L. P., Sorbring, E., Al-Hassan, S. M., Oburu, P., Malone, P. S., Di Giunta, L., Tirado, L. M. U., & Tapanya, S. (2019). Environmental harshness and unpredictability, life history, and social and academic behavior of adolescents in nine countries. *Developmental Psychology*, *55*(4), 890–903. https://doi.org/10.1037/dev0000655

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/s15328007sem0902_5

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Coldwell, J., Pike, A., & Dunn, J. (2006). Household chaos: Links with parenting and child behaviour. *Journal of Child Psychology and Psychiatry*, *47*(11), 1116–1122. https://doi.org/10.1111/j.1469-7610.2006.01655.x

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. https://doi.org/10.1037/a0033805

Deater-Deckard, K., Godwin, J., Lansford, J. E., Tirado, L. M. U., Yotanyamaneewong, S., Alampay, L. P., Al-Hassan, S. M., Bacchini, D., Bornstein, M. H., Chang, L., Di Giunta, L., Dodge, K. A., Oburu, P., Pastorelli, C., Skinner, A. T., Sorbring, E., Steinberg, L., & Tapanya, S. (2019). Chaos, danger, and maternal parenting in families: Links with adolescent adjustment in low- and middle-income countries. *Developmental Science*, *22*(5), e12855. https://doi.org/10.1111/desc.12855

Deater-Deckard, K., Mullineaux, P. Y., Beekman, C., Petrill, S. A., Schatschneider, C., & Thompson, L. A. (2009). Conduct problems, IQ, and household chaos: A longitudinal multi-informant study. *Journal of Child Psychology and Psychiatry*, *50*(10), 1301–1308. https://doi.org/10.1111/j.1469-7610.2009.02108.x

Dewald, J. F., Meijer, A. M., Oort, F. J., Kerkhof, G. A., & Bögels, S. M. (2010). The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: A meta-analytic review. *Sleep Medicine Reviews*, *14*(3), 179–189. https://doi.org/10.1016/j.smrv.2009.10.004

Dumas, J. E., Nissley, J., Nordstrom, A., Smith, E. P., Prinz, R. J., & Levine, D. W. (2005). Home chaos: Sociodemographic, parenting, interactional, and child correlates. *Journal of Clinical Child & Adolescent Psychology*, *34*(1), 93–104. https://doi.org/10.1207/s15374424jccp3401_9

Dunn, L. M., & Markwardt, F. C. (1998). *Peabody Individual Achievement Test—Revised/Normative Update*. American Guidance Service.

Elardo, R., Bradley, R., & Caldwell, B. M. (1977). A longitudinal study of the relation of infants' home environments to language development at age three. *Child Development*, *48*(2), 595–603. https://doi.org/10.2307/1128658

Evans, G. W. (2006). Child development and the physical environment. *Annual Review of Psychology*, *57*(1), 423–451. https://doi.org/10.1146/annurev.psych.57.102904.190057

Flake, J. K. (2021). Strengthening the foundation of educational psychology by integrating construct validation into open science reform. *Educational Psychologist*, *56*(2), 132–141. https://doi.org/10.1080/00461520.2021.1898962

Ganasegeran, K., Selvaraj, K., & Rashid, A. (2017). Confirmatory factor analysis of the Malay version of the Confusion, Hubbub and Order Scale (CHAOS-6) among myocardial infarction survivors in a Malaysian cardiac healthcare facility. *Malaysian Journal of Medical Sciences*, *24*(4), 39–46. https://doi.org/10.21315/mjms2017.24.4.5

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or 'p-hacking" and the research hypothesis was posited ahead of time.* http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gould, K. L., Coventry, W. L., Olson, R. K., & Byrne, B. (2018). Gene-environment interactions in ADHD: The roles of SES and chaos. *Journal of Abnormal Child Psychology*, *46*(2), 251–263. https://doi.org/10.1007/s10802-017-0268-7

Graafland, J. H. (2018). New technologies and 21st century children: Recent trends and outcomes. *OECD Education Working Papers*. https://doi.org/10.1787/e071a505-en

Hancock, G. R. (2003). Fortune cookies, measurement error, and experimental design. *Journal of Modern Applied Statistical Methods*, *2*(2), 293–305. https://doi.org/10.22237/jmasm/1067644980

Hanscombe, K. B., Haworth, C. M. A., Davis, O. S. P., Jaffee, S. R., & Plomin, R. (2010). The nature (and nurture) of children's perceptions of family chaos. *Learning and Individual Differences*, *20*(5), 549–553. https://doi.org/10.1016/j.lindif.2010.06.005

Hanscombe, K. B., Haworth, C. M. A., Davis, O. S. P., Jaffee, S. R., & Plomin, R. (2011). Chaotic homes and school achievement: A twin study. *Journal of Child Psychology and Psychiatry*, *52*(11), 1212–1220. https://doi.org/10.1111/j.1469-7610.2011.02421.x

Harlaar, N., Butcher, L. M., Meaburn, E., Sham, P., Craig, I. W., & Plomin, R. (2005). A behavioural genomic analysis of DNA markers associated with general cognitive ability in 7-year-olds. *Journal of Child Psychology and Psychiatry*, *46*(10), 1097–1107. https://doi.org/10.1111/j.1469-7610.2005.01515.x

Hart, S. A., Petrill, S. A., Deater-Deckard, K., & Thompson, L. A. (2007). SES and CHAOS as environmental mediators of cognitive ability: A longitudinal genetic analysis. *Intelligence*, *35*(3), 233–242. https://doi.org/10.1016/j.intell.2006.08.004

Hart, S. A., Schatschneider, C., Reynolds, T. R., Calvo, F. E., Brown, B. J., Arsenault, B., Hall, M. R. K., van Dijk, W., Edwards, A. A., Shero, J. A., Smart, R., & Phillips, J. S. (2020). *LDBase: A Learning and Development Data Repository*. https://doi.org/10.33009/ldbase

Johnson, A. D., Martin, A., Brooks-Gunn, J., & Petrill, S. A. (2008). Order in the house! Associations among household chaos, the home literacy environment, maternal reading ability, and children's early reading. *Merrill-Palmer Quarterly*, *54*(4), 445–472. https://doi.org/10.1353/mpq.0.0009

Kenny, D. (1979). *Correlation and Causality*. Wiley.

Kim-Spoon, J., Maciejewski, D., Lee, J., Deater-Deckard, K., & King-Casas, B. (2017). Longitudinal associations among family environment, neural cognitive control, and social competence among adolescents. *Developmental Cognitive Neuroscience*, *26*, 69–76. https://doi.org/10.1016/j.dcn.2017.04.009

Kings College London. (2022). *Twins Early Development Study*. https://www.teds.ac.uk/

Larsen, S. A., Little, C. W., Grasby, K., Byrne, B., Olson, R. K., & Coventry, W. L. (2020). The academic development study of Australian twins (ADSAT): Research aims and design. *Twin Research and Human Genetics*, *23*(3), 165–173. https://doi.org/10.1017/thg.2020.49

Lauharatanahirun, N., Maciejewski, D., Holmes, C., Deater-Deckard, K., Kim-Spoon, J., & King-Casas, B. (2018). Neural correlates of risk processing among adolescents: Influences of parental monitoring and household chaos. *Child Development*, *89*(3), 784–796. https://doi.org/10.1111/cdev.13036

Laurent, H. K., Neiderhiser, J. M., Natsuaki, M. N., Shaw, D. S., Fisher, P. A., Reiss, D., & Leve, L. D. (2014). Stress system development from age 4.5 to 6: Family environment predictors and adjustment implications of HPA activity stability versus change. *Developmental Psychobiology*, *56*(3), 340–354. https://doi.org/10.1002/dev.21103

Little, C. W., Hart, S. A., Phillips, B. M., Schatschneider, C., & Taylor, J. E. (2019). Exploring neighborhood environmental influences on reading comprehension. *Journal of Applied Developmental Psychology*, *62*, 173–184. https://doi.org/10.1016/j.appdev.2019.02.009

Marsh, S., Dobson, R., & Maddison, R. (2020). The relationship between household chaos and child, parent, and family outcomes: A systematic scoping review. *BMC Public Health*, *20*(1), 513. https://doi.org/10.1186/s12889-020-08587-8

Matheny, A. P., Jr., Wachs, T. D., Ludwig, J. L., & Phillips, K. (1995). Bringing order out of chaos: Psychometric characteristics of the confusion, hubbub, and order scale. *Journal of Applied Developmental Psychology*, *16*(3), 429–444. https://doi.org/10.1016/0193-3973(95)90028-4

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(Suppl 3), S69–S77. https://doi.org/10.1097/01.mlr.0000245438.73837.89

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 380–392). The Guilford Press.

Oliver, B. R., & Plomin, R. (2007). Twins' Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Research and Human Genetics*, *10*(1), 96–105. https://doi.org/10.1375/twin.10.1.96

Organisation for Economic Cooperation and Development. (2010). *TALIS 2008 Technical Report*. https://www.oecd.org/education/school/44978960.pdf

Petrill, S. A., Deater-Deckard, K., Thompson, L. A., De Thorne, L. S., & Schatschneider, C. (2006). Reading skills in early readers: Genetic and shared environmental influences. *Journal of Learning Disabilities*, *39*(1), 48–55. https://doi.org/10.1177/00222194060390010501

Petrill, S. A., Pike, A., Price, T., & Plomin, R. (2004). Chaos in the home and socioeconomic status are associated with cognitive development in early childhood: Environmental mediators identified in a genetic design. *Intelligence*, *32*(5), 445–460. https://doi.org/10.1016/j.intell.2004.06.010

Peviani, K. M., Kahn, R. E., Maciejewski, D., Bickel, W. K., Deater-Deckard, K., King-Casas, B., & Kim-Spoon, J. (2019). Intergenerational transmission of delay discounting: The mediating role of household chaos. *Journal of Adolescence*, *72*(1), 83–90. https://doi.org/10.1016/j.adolescence.2019.03.002

Pike, A., Iervolino, A. C., Eley, T. C., Price, T. S., & Plomin, R. (2006). Environmental risk and young children's cognitive and behavioral development. *International Journal of Behavioral Development*, *30*(1), 55–66. https://doi.org/10.1177/0165025406062124

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2013). *Behavioral Genetics* (6th ed.). Worth Publisher.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Revelle, W. (n.d.). Classical test theory and the measurement of reliability. In *An introduction to psychometric theory with applications in R* (pp. 205–239). https://personality-project.org/r/book/#

Revelle, W. (2022). *Psych: Procedures for psychological, psychometric, and personality research*. http://personality-project.org/r/psych/psych-manual.pdf

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Rimfeld, K., Malanchini, M., Spargo, T., Spickernell, G., Selzam, S., McMillan, A., Dale, P. S., Eley, T. C., & Plomin, R. (2019). Twins early development study: A genetically sensitive investigation into behavioral and cognitive development from infancy to emerging adulthood. *Twin Research and Human Genetics*, *22*(6), 508–513. https://doi.org/10.1017/thg.2019.56

Rosseel, Y. (2012). lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*(7781), 9. https://doi.org/10.1038/d41586-019-03350-5

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*(1), 102–111. https://doi.org/10.1037/1040-3590.12.1.102

Suku, S., Soni, J., Martin, M. A., Mirza, M. P., Glassgow, A. E., Gerges, M., Van Voorhees, B. W., & Caskey, R. (2019). A multivariable analysis of childhood psychosocial behaviour and household functionality. *Child: Care, Health and Development*, *45*(4), 551–558. https://doi.org/10.1111/cch.12665

Taylor, J., Martinez, K., & Hart, S. A. (2019). The Florida State Twin Registry. *Twin Research and Human Genetics*, *22*(6), 728–730. https://doi.org/10.1017/thg.2019.102

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

van Dijk, W., Norris, C. U., Otaiba, S. A., Schatschneider, C., & Hart, S. A. (2022). Exploring individual differences in response to reading intervention: Data from project KIDS (Kids and Individual Differences in Schools). *Journal of Open Psychology Data*, *10*(1), 2. https://doi.org/10.5334/jopd.58

van Dijk, W., Schatschneider, C., & Hart, S. A. (2021). Open science in education sciences. *Journal of Learning Disabilities*, *54*(2), 139–152. https://doi.org/10.1177/0022219420945267

von Stumm, S., Starr, A., Voronin, I., & Malanchini, M. (2023, February 8). *"Chaos is the score upon which reality is written": A genetically informative study on the developmental interplay between household chaos and educational achievement from age 9 through 16 years*. https://doi.org/10.31234/osf.io/z5dqf

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 209–231). The Guilford Press.

Wickham, H. (2014). *ggplot2: An implementation of the grammar of graphics*. https://cran.microsoft.com/snapshot/2015-01-06/web/packages/ggplot2/ggplot2.pdf

Widaman, K. F., & Revelle, W. (2022). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, *55*(2), 788–806. https://doi.org/10.3758/s13428-022-01849-w

Wilson, R. S., & Matheny, A. P., Jr. (1983). Mental development: Family environment and genetic influences. *Intelligence*, *7*(2), 195–215. https://doi.org/10.1016/0160-2896(83)90029-6

Woodcock, R. W. (1987). *Woodcock reading mastery tests-revised*. American Guidance Service.

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/77837-measuring-chaos-evaluating-the-short-form-confusion-hubbub-and-order-scale/attachment/163234.docx?auth_token=cYxjg7Wgnpx_Nvi-NoxB

Supplemental Materials

Download: https://collabra.scholasticahq.com/article/77837-measuring-chaos-evaluating-the-short-form-confusion-hubbub-and-order-scale/attachment/163235.docx?auth_token=cYxjg7Wgnpx_Nvi-NoxB