International Journal of Methods in Psychiatric Research
14(4): 186–201 (2005)
DOI: 10.1002/mpr.7

ADHD Rating Scale IV: psychometric properties from a multinational study as a clinician-administered instrument

S. ZHANG,[1] D.E. FARIES,[1] M. VOWLES,[2] D. MICHELSON[1,3]

1 Lilly Research Laboratories, Indianapolis, USA
2 Lilly Research Centre, Surrey, UK
3 Indiana University School of Medicine, Indianapolis, USA

ABSTRACT

The development of rating scales for attention-deficit/hyperactivity disorder (ADHD) has traditionally focused on parent- or teacher-rated scales. However, clinician-based instruments are valuable tools for assessing ADHD symptom severity. The ADHD Rating Scale IV (ADHD RS), clinician administered and scored, has been validated as a useful instrument to assess ADHD symptoms among American children and adolescents. In this study, we assessed the psychometric properties of the scale in a recent clinical trial conducted mainly in Europe with over 600 children and adolescents diagnosed with ADHD. The trial was conducted in 11 European countries plus Australia, Israel, and South Africa.

Results based on data in the study indicate that this version of the scale has acceptable psychometric properties including inter-rater reliability, test-retest reliability, internal consistency, factor structure, convergent and divergent validity, discriminant validity, and responsiveness. There were low-to-moderate ceiling and floor effects. The psychometric properties were comparable with other validated scales for assessing ADHD symptom severity. These results were consistent across the 14 countries participating in this trial. Overall, the data from this study support the use of the ADHD RS as a clinician-rated instrument for assessing the severity of ADHD symptoms in children and adolescents in Europe. Copyright © 2005 John Wiley & Sons, Ltd.

Key words: Attention Deficit Hyperactivity Disorder, Rating Scale-IV-Parent, psychometric properties

Information about the scale rating procedure is given in the method section of this paper.

Introduction

Attention-deficit/hyperactivity disorder (ADHD) is a neurobiological disorder characterized by symptoms of inattention and/or hyperactivity/impulsivity with an onset during childhood (Pliszka et al., 1996; Faraone and Biederman, 1998). It is one of the most common psychiatric disorders of childhood and adolescence, and occurs in 3% to 7% of school-aged children in the US (American Psychiatric Association, 2000). Outside of the US and Canada, the reported prevalence rates for ADHD vary from country to country (Mueller et al., 1995; Reid et al., 1998; Livingston, 1999; Graetz at al., 2001). In the UK, the estimated prevalence rate of ADHD has been reported to be 5% for school-aged children (National Institute of Clinical Excellence, 2000). In the Netherlands, a prevalence of 2% to 4% for children between the ages of 5 and 14 years has been reported (Health Council of the Netherlands, 2000), while Baumgaertel, Wolraich, and Dietrich (1995) found a prevalence rate of 17.8% for attention deficit disorders in German school-aged children. These differences, however, appear to be primarily due to different diagnostic criteria and methodology rather than true differences in prevalence, as evidenced when consistent criteria are used to assess the presence of the disorder (Prendergast et al., 1988).

The standard instruments for assessing ADHD symptom severity have traditionally been parent- or

teacher-rated scales, which have been well validated and in use for many years (Barkley, 1990; Conners, 1997; DuPaul et al., 1998). Cohen et al. (1990) concluded that the low correlation between parent- and teacher-scored scales was not due to low reliability of the scales but suggested the need for multiple methods for assessing symptom severity. Brown et al. (2001) found that the teacher rating scale had a lower effect size when compared with the parent- and clinician-rated scales. Recently, Beiderman et al. (2004) performed a literature search and summarized clinical trials, which contained both parent- and teacher-reported measures. Their results showed similar sensitivity for parent and teacher ratings.

Parent- and teacher-rated scales provide important information for assessing ADHD symptom severity but a clinician-rated assessment may be preferable as a primary outcome measure in a clinical trial designed to assess treatment efficacy. First of all, clinicians are trained to assess the impact of interventions already implemented to cope with ongoing behaviour problems (for example, special classroom methods), an important factor in achieving an accurate assessment of the severity of the underlying disorder. Secondly, use of a clinician-rated scale avoids the problems of needing to obtain rating for children from multiple teachers and issues of school vacations. Thirdly, inclusion of clinician assessments is consistent with the DSM-IV criteria for the diagnosis of ADHD, which requires that the ADHD symptoms are present to a degree of impairment in at least 2 settings, which cannot be satisfied by parent or teacher ratings alone. Furthermore, in the clinical trial setting, trained clinician raters can apply standardized, symptom-severity inclusion criteria, which lead to the selection of a more homogenous patient population and a reduction in variability that can be important in detecting true drug effects. Additionally, application of standardized inclusion criteria (DSM-IV) by the trained clinicians facilitates the incorporation of data collected from multiple sources and settings into a single score, thereby reducing the statistical multiplicity that occurs when data are collected inconsistently using separate measures. And finally, in neuroscience clinical trials, using trained clinician raters can reduce variability by having them apply consistent judgements on severity ratings across patients; therefore, clinician-rated instruments of illness severity are typically required as the primary efficacy outcome measures by regulatory authorities.

Parent-, teacher-, and clinician-scored versions of the ADHD Rating Scale-IV (ADHD RS, DuPaul et al., 1998; Faries, 2001) have been validated in North American populations. Magnusson et al. (1999) studied the validity of an Icelandic version of ADHD RS rated by parents and teachers in over 400 Icelandic schoolchildren. However, the validity of this instrument has not been previously studied in majority non-North American populations, especially as a clinician-rated instrument.

The objective is to examine the psychometric properties of the ADHDRS-IV-Parent: Clinician administrated and scored (ADHDRS-PI), scored by trained clinicians based on parent interviews in children with a diagnosis of ADHD who reside outside of North America. Specifically, the assessment includes the inter-rater reliability, internal consistency, factor structure, test-retest reliability, convergent and divergent validity, discriminant validity, responsiveness, and ceiling and floor effects. These psychometric properties are assessed relative to Clinical Global Impressions-ADHD-Severity (CGI-ADHD-S), Conners' Parent Rating Scales (CPRS), and Conners' Teacher Rating Scales (CTRS).

Materials and methods

Study design and rating scale

The study of the validity and reliability of the ADHDRS-PI was assessed as part of a large clinical trial of atomoxetine. The study was conducted at 33 sites in the UK, France, Spain, Italy, Belgium, the Netherlands, Germany, Poland, Hungary, Sweden, Norway, Israel, South Africa, and Australia, and enrollment took place over approximately 11 months. Principle investigators at each site were physicians with specialty training in psychiatry or paediatrics and psychologists with experience diagnosing and treating children and adolescents with ADHD. After description of the procedures, purpose of the study, and prior to the administration of any study procedure or dispensing of study medication, written informed consent was obtained from each patient's parent or guardian and written assent was obtained from each patient. This study was reviewed by each site's institutional review board and was conducted in accordance with the ethical standards of the Declaration of Helsinki 1975, as revised in 2000.

In this trial, 604 children and adolescents, aged 6 through 15 years, who met the Diagnostic and

Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria for a diagnosis of ADHD, were enrolled at 33 investigational sites in 14 countries. The study design included a 1-week assessment and drug washout phase for those children and adolescents taking any medication excluded by the protocol, followed by a 10-week, open-label, treatment period during which atomoxetine was administrated twice a day (the subsequent extension phase of the study is ongoing). Detailed information regarding the efficacy and safety of the use of atomoxetine has been previously published (Allen et al., 2001; Michelson et al., 2001; Spencer et al., 2001; Michelson et al., in press). Only data from the acute phase are presented here.

Patients were children and adolescents 6 through 15 years of age at the time of study entry. For each patient, the clinical diagnosis of ADHD was confirmed using the Kiddie Schedule of Affective Disorders and Schizophrenia for School-Aged Children Present and Lifetime Version (KSADS-PL) (Kaufman et al., 1996). Patients must have had an ADHDRS-PI score of at least 1.5 standard deviations above the age and gender norm for their diagnostic subtype using published US norms for the ADHDRS-PI. Generally, the same investigator completed both KSADS-PL and the ADHD RS. Patients who were taking psychotropic medication at study entry had a washout equal to a minimum of five half-lives of the psychotropic medication prior to obtaining the baseline severity assessment and starting treatment with atomoxetine. Patients were seen at approximately weekly or biweekly visits during the 10-week, open-label period, and no other psychotropic medications were allowed during the study.

Parent-, teacher-, and clinician-scored instruments were used in this study to assess the severity of ADHD. Since clinicians at each investigational site were fluent in English, the English version of clinician-rated scales was used without translation into the native languages. The intention of this paper is simply validating the use of the English clinician-rated ADHDRS-PI scale in countries outside the US. The translation and validation of self and investigator forms is the subject of future work.

The primary efficacy measure was the ADHDRS-PI. It was completed at each visit to assess ADHD symptom severity over the past week or past 2 weeks. The ADHDRS-PI is an 18-item scale with one item for each of the 18 symptoms contained in the DSM-IV diagnosis of ADHD. Each item is scored on a 0 to 3 scale: 0 = none (never or rarely); 1 = mild (sometimes); 2 = moderate (often); 3 = severe (very often). The total score is computed as the sum of the scores on each of the 18 items. In addition to the total score, the scores from the inattention and hyperactivity/impulsivity subscales were computed. The inattention subscale score is the sum of the scores on the odd-numbered items and the hyperactivity/impulsivity subscale is the sum of the scores on the even-numbered items. A clinician-rated symptom severity for each item was based on his or her interview with the child's parent or primary caretaker. The clinicians were trained at the time of study start-up to rate scores based on the frequency of the behaviour (across multiple settings) and the degree of impairment, and to use developmental comparison (to consider the age appropriateness of behaviour in rating each item).

The Clinical Global Impressions-ADHD-Severity (CGI-ADHD-S) (Guy, 1976) is a single-item clinician rating of the clinician's assessment of the severity of the ADHD symptoms in relation to the clinician's total experience with patients with ADHD. Severity is rated on a 7-point scale: 1 = normal, not ill; 2 = minimally ill; 3 = mildly ill; 4 = moderately ill; 5 = markedly ill; 6 = severely ill; and 7 = very severely ill. It was completed at each visit. Other scales used in this study included the Conners' Parent Rating Scale – Revised: Short Form (CPRS) and Conners' Teacher Rating Scale-Revised: Short Form (CTRS) (Conners, 1997). Both the CPRS and CTRS have 4 subscales and were collected at Visit 1 and endpoint: ADHD Index, Hyperactivity, Cognitive, and Oppositional.

Reliability and validity methods

Inter-rater reliability

Inter-rater reliability refers to a scale's ability to achieve similar ratings by different raters assessing the same patient. The Kappa statistic and average squared deviation from the mode were used to assess inter-rater reliability. Kappa statistic has a range from 0 to 1 and is commonly used to assess inter-rater reliability when observing categorical variables. The Kappa statistic is the proportion of agreeing pairs out of all possible pairs, adjusted for chance agreement (Fleiss, 1971). The average squared deviation was computed for each rater by averaging (over the 18 items) the squared

difference between their ratings and the mode rating for each item from the entire group (Channon and Butler, 1998); values from all raters were then averaged. Unlike the Kappa statistic, the average squared deviation statistic more severely penalizes ratings that are more than 1 point from the mode.

For this study, all clinical personnel who would be using the ADHDRS-PI were trained and assessed in 2 rater-training sessions prior to the study. The rater-training session began with a presentation and discussion of the rating scale. Raters then watched a videotaped interview between a clinician and a parent of a child with ADHD, and independently completed the ADHDRS-PI after the video. The results were then presented and points of agreement and disagreement were discussed to help gain consistency in future scoring. The raters then watched a second videotaped interview and completed the ADHDRS-PI for the second time. Data from 1 of the 2 rater-training sessions were collected and available for the analysis in this manuscript.

Factor structure

Factor analysis is a powerful tool used to uncover the latent structure (dimensions) of a set of variables. It can be used to validate a scale by demonstrating that its constituent items load on the same factor, and to drop proposed scale items that cross-load on more than 1 factor.

In this paper, a principal-components factor analysis with Varimax rotation (Reid, 1995) was performed to examine the factor structure of ADHD RS.

Internal consistency

Cronbach's alpha was used to evaluate internal consistency (Cronbach, 1951). It assesses the degree to which each item of a rating scale measures the same construct based upon all possible correlations between 2 sets of items within a rating scale. The range of the statistic is from 0 to 1. The accepted minimal standard to claim internal consistency is 0.65 (Nunnally, 1994). The scores for all 18 items in ADHDRS-PI at Visit 1 for all enrolled patients were used to compute Cronbach's alpha for ADHDRS-PI total score. Scores from all odd- and even-numbered items were used to compute Cronbach's alpha for ADHDRS-PI inattention subscale and hyperactive/impulsive subscale. Similar analysis was also done for ADHDRS-PI at Visit 2.

Test-retest reliability

A good rating scale should be able to reproduce the same score for the same individual at different times while in the same disease condition. Test-retest reliability indicates such stability of a rating scale.

The intra-class correlation coefficient (ICC) is the recommended measure to assess the test-retest reliability (Deyo et al., 1991). The ICC is computed as the variability due to patients divided by the total variability (included variability due to patients, time, and other factors). Since Pearson's correlation coefficient is the most commonly used measure for assessing the strength of the relationship between the scores, both coefficients were computed to assess the correlation between the ADHD-PI total scores at Visits 1 and 2. To show the variability of correlation, the confidence intervals for ICC and Pearson's correlation coefficients were computed based on Fisher's Z transformation (Anderson, 1990). To further quantify shifts over time, t-test was used to test the null hypothesis that the mean ADHD total scores obtained were similar at Visits 1 and 2.

In this study, ADHDRS-PI and CGI-ADHD-S were completed at both Visits 1 and 2, which were scheduled approximately 1 week apart. Since no study drug was dispensed at these two visits, the changes in ADHDRS-PI and CGI-ADHD-S scores from Visits 1 to 2 can be used to assess the stability of scores. Patients who were taking medication for ADHD other than the study drug at Visit 1 were excluded from the analysis group due to the possible changes in ADHD symptoms from Visits 1 to 2 after discontinuing the drug. Test-retest reliability was not assessed for the CPRS and CTRS as these scales were only administered at Visit 1.

In addition to the correlation analysis, a graphical approach, Bland and Altman Plot (Bland and Altman, 1986), which plots the difference between the 2 measures plot against the average score, was also used to assess the test-retest reliability.

Starting from this section, the imputed scores for total and subscales are used for all analyses. The imputed score for subscale was computed as follows: if only 1 single item was missing in the subscale, the mean score for all other items in the subscale was imputed as the score for the missing score. The total score was computed as the sum of the imputed subscale scores. If more than 1 item was missing in the subscale then the subscale score and total score would be missing.

Convergent and divergent validity

Convergent and divergent validity were utilized to establish the construct validity of the scale. Convergent/divergent validity indicates a relationship between the scale under review and other scales thought to measure the same/different construct. In this manuscript, the scale under review is ADHDRS-PI; other validated scales for comparison were the CPRS (parent scored), CTRS (teacher scored), and CGI-ADHD-S (investigator scored). All of these scales were used to measure the severity of ADHD symptoms.

To assess convergent validity, Pearson's correlation coefficients were computed between the following measures (measures of the same construct): ADHDRS-PI Total with CGI-ADHD-S, CPRS ADHD Index and CTRS ADHD Index, ADHDRS-PI Inattention Subscale with CPRS Cognitive and CTRS Cognitive Subscale, and ADHDRS-PI hyperactive/impulsive subscale with CPRS hyperactive and CTRS hyperactive subscale. Correlations were computed for both baseline and change-from-baseline-to-endpoint scores.

To assess divergent validity, Pearson's correlation coefficients were computed between the following measures (measures of different construct): ADHDRS-PI Inattention Subscale with CPRS hyperactive and CTRS hyperactive subscale, and ADHDRS-PI hyperactive/impulsive subscale with CPRS cognitive and CTRS cognitive subscale. Correlations were computed for baseline scores.

To show the variability of correlation, the confidence intervals for Pearson's correlation coefficients were computed based on Fisher's Z transformation (Anderson, 1990).

Discriminant validity

The discriminant validity of a scale measures the scales ability to distinguish between different groups of subjects. An instrument for assessing ADHD symptom severity with discriminant validity should distinguish between patients with and without a diagnosis of ADHD, or between patients with and without significant hyperactive symptoms. In this study, only patients with ADHD symptoms were recruited. Thus, instead of a comparison between ADHD patients and a control group, 3 alternative approaches were used to assess the discriminant validity.

First, a comparison between patients with different ADHD subtypes, namely inattentive, hyperactive/impulsive, and combined (inattentive plus hyperactive /impulsive) was conducted. As noted previously, ADHD subtype was assessed at Visit 1 using the KSADS-PL. Patients with an inattentive subtype did not have sufficient hyperactive/impulsive symptomatology to meet the full combined ADHD subtype. This provides an opportunity to assess whether the hyperactive/impulsive subscale of the ADHDRS-PI was able to distinguish between the ADHD subtypes. Therefore, we think the comparison between subtypes can be used to assess the discriminant validity of a subscale of the ADHDRS-PI.

Secondly, although patients without ADHD symptoms were not recruited, with active treatment, severity levels of the symptoms decrease to about normal (as indicated by CGI-ADHD-S score = 1 or 2) at endpoint for some patients. Thus, an analysis of variance can be used to compare mean ADHDRS-PI total scores with CGI-ADHD-S scores. Comparison with the CGI-ADHD-S scores also provides additional information regarding the clinical significance of specific ADHDRS-PI total scores.

Third, Pearson's correlation coefficient between the ADHDRS-PI and other measures that should not be logically related were computed. Several measurements included in the study were the Children's Depression Inventory (CDI) and Children's Depression Rating Scale-Revised (CDRS) to measure presence and severity of depression, and the Multidimensional Anxiety Scale for Children (MASC) to assess anxiety. The correlation of scores between ADHDRS-PI and these measurements should be low compared with that between ADHDRS-PI and CPRS or CTRS.

Responsiveness

Responsiveness indicates the ability of a scale to detect small but clinically significant changes in the patient's symptom severity when a change has occurred. The standardized response mean (SRM) is a commonly used statistic to assess responsiveness (Stratford et al., 1996). The SRM is defined as the mean change-from-baseline score divided by the standard deviation of the changes scores. The SRM based on the ADHDRS-PI was compared with the SRM from other validated scales (CGI-ADHD-S, CPRS, and CTRS).

Minimal clinically important differences

Another important need is to determine the between- and within-treatment minimum clinically important

differences (MCID) for an instrument. The MCID helps clinicians interpret the relevance of changes in the instrument scores. The within-treatment MCID is defined as the improvement in a score with treatment at which a patient recognizes that she/he is improved. The between-treatment MCID is the minimum difference between 2 treatments that can be considered clinically relevant. One widely accepted way to determine the MCIDs is to anchor the scale to a global rating scale such as CGI-Improvement (CGI-I). Unfortunately, CGI-I was not collected in this study; as an alternative, CGI-ADHD-S was used. First, the LOCF change from baseline to endpoint of CGI-ADHD-S was calculated for each patient. If the change score is 0, then the patient was rated as having 'no change'; if the change score is 1 (the smallest change detectable by the CGI-ADHD-S), then the patient was rated as 'a little better'. The mean change in the ADHDRS for those subjects who rated as 'a little better' could be considered as the within-treatment MCID. The difference in the mean changes for subjects who rated as 'a little better' and who rated 'no change' could be considered as the between-treatment MCID. The between-treatment MCID can be a sound choice for the treatment difference in order to power the clinical studies. These two MCIDs provide guidance to researchers to interpret the change scores for the instrument. They become critical when statistically significant differences needed to be justified as clinically relevant.

Ceiling and floor effects

Ceiling and floor effects exist if a substantial percentage of the patient scores are at the ends of the scales – then the scale may not be able to accurately capture change scores or differentiate among patients near the ceiling or floor. A scale with floor effect lacks the ability to detect minor disease symptoms, while a scale with ceiling effect would be less sensitive to changes in the more serious symptoms (Stucki and Michel, 1995; Herrmann et al., 1997). Percentages of lowest and highest possible scores at baseline and endpoint were calculated to assess ceiling and floor effects.

Results

Subjects

Six-hundred-and-four patients enrolled in this study (14 countries, 33 investigational sites). Table 1 summarizes the patient characteristics for this group as well as the patient characteristics for a similar US-based study. The mean (SD) age for the group was 10.24 (2.25) years.

Table 1. Summary of baseline patient characteristics

	Current multinational study (N=604)	Previous US study (Faries, 2001) (N=228)
Variable	n (%)	n (%)
Gender		
Male	541 (89.6)	211 (92.5)
Origin		
Caucasian	583 (96.5)	175 (76.8)
ADHD subtype		
Hyper/impulsive	30 (5.0)	3 (1.6)
Inattentive	124 (20.5)	52 (22.9)
Combined	450 (74.5)	172 (75.8)
Previous stimulant	341 (56.5)	125 (54.8)
Age (yrs.)		
5–7	44 (7.3)	
8–9	155 (25.7)	69 (30.3)
10–11	184 (30.5)	72 (31.6)
12–13	140 (23.2)	51 (22.4)
14–15	81 (13.4)	36 (15.8)

The diagnosis of ADHD and comorbid diagnoses were assessed by clinical interview and confirmed using the KSADS-PL semi-structured interview. In addition to the diagnosis of ADHD, 45.5% of the patients also had a diagnosis of oppositional defiant disorder, 5.6% had conduct disorder, and 1.5% had depression.

Inter-rater reliability

Rater training data from 41 raters were collected for analysis prior to the start of the trial. The Kappa statistics for the first and second tapes were 0.58 and 0.63, respectively. The average squared deviations from the mode for the first and second tapes for the rating scale were 0.38 and 0.30, respectively. Agreement was similar for inattention and hyperactive/impulsive items. Approximately 80% of the raters independently gave a total score in the range of 33 through 39.

Factor structure

The first three eigenvalues in the solution were 5.99 (55%), 2.80 (26%), and 0.80 (7%). Kaiser's criterion of an eigenvalue greater than 1 indicated that two factors could be extracted for ADHDRS-PI. Table 2 shows the structure matrix for the two-factor solution. The pattern of loadings also showed a clear hyperactivity-impulsivity factor and a clear inattention factor. Odd-numbered items (reflecting inattention) loaded on Factor 2 and even-numbered items (reflecting hyperactivity) loaded on Factor 1.

Internal consistency

Cronbach's alpha for the ADHDRS-PI total score was 0.795 based on Visit 1 data and 0.838 based on Visit 2 data from the 604 enrolled patients (for the inattention subscale, 0.724 for Visit 1 and 0.770 for Visit 2; for the hyperactivity-impulsivity subscale, 0.825 for Visit 1 and 0.848 for Visit 2). The item-to-total correlations range from 0.25 to 0.51 at Visit 1, and from 0.29 to 0.57 at Visit 2.

Test-retest reliability

Test-retest assessments included the 565 patients who were not taking stimulant medication for ADHD at Visit 1 and who had an efficacy measurement (ADHDRS-PI) at both Visits 1 and 2. Table 3 reports the ICC, the Pearson's correlation coefficients, the mean changes and their associated confidence intervals for ADHDRS-PI total score, ADHDRS-PI inattentive subscale score, ADHDRS-PI hyperactive/

Table 2. Factor pattern matrix of a Varimax rotation for the ADHD RS

Item #	Item short description	Factor 1	Factor 2	Communality
Hyperactivity				
10	On the go	0.71	0.02	0.50
06	Runs about	0.68	0.09	0.48
16	Difficult waiting turn	0.63	0.09	0.41
18	Interrupts	0.62	0.08	0.39
08	Difficult playing	0.61	0.16	0.39
14	Blurts out answers	0.54	0.13	0.31
04	Leaves seat	0.53	0.06	0.29
12	Talks excessively	0.51	0.04	0.26
02	Fidgets	0.46	0.12	0.22
Inattentive				
07	No follow-through	0.04	0.60	0.36
17	Forgetful	0.06	0.57	0.33
09	Difficult organizing	0.00	0.56	0.31
11	Avoids tasks	0.04	0.54	0.29
01	Close attention	0.05	0.52	0.27
15	Easily distracted	0.13	0.48	0.25
03	Sustaining attention	0.14	0.44	0.22
13	Loses things	0.10	0.42	0.19
05	Does not listen	0.27	0.35	0.19

Table 3. Test-retest reliability of ADHDRS-PI and CGI-ADHD-S

Variable	*N*	ICC (LCI, UCI)*	Pearson's correlation (LCI, UCI)**	Mean difference (LCI, UCI)***
ADHDRS-PI Total score	565	.84 (.82, .87)	.85 (.82, .87)	.09 (–.25, .44)
ADHDRS-PI Hyper/imp Subscale	565	.89 (.87, .90)	.89 (.87, .90)	.19 (–.03, .40)
ADHDRS-PI Inattentive Subscale	565	.77 (.74, .80)	.78 (.74, .81)	–.09 (–.31, .12)
CGI-ADHD-S score	566	.85 (.83, .87)	.86 (.83, .88)	.02 (–.01, .06)

ICC: Intra-class coefficient from ANOVA model with term of PATIENT and VISIT; mean difference: mean of change score from Visits 1 to 2.
LCI – lower bound of 95% confidence interval.
UCI – upper bound of 95% confidence interval.
* 95% confidence interval of ICC based on Fisher's Z transformation.
** 95% confidence interval of Pearson's correlation between Visit 1 and Visit 2 scores. Based on Fisher's Z transformation.
*** 95% confidence interval of mean difference based on normal distribution.

impulsive subscale score, and CGI-ADHD-S score. The correlation coefficients for the ADHDRS-PI scale and the CGI-ADHD-S scale were both high (range from 0.78 to 0.89), and the mean differences in scores from Visits 1 to 2 were both very low (range from –0.09 to 0.19).

Figure 1 plots the difference of ADHD RS total score between Visits 1 and 2 against the average of ADHDRS-PI total scores at those 2 visits. The mean difference was –0.14, and the limits of agreement were –8.9 to 8.6.

Convergent and divergent validity

Pearson's correlation coefficients between ADHDRS-PI scores and scores from other scales thought to measure the same construct for both baseline and

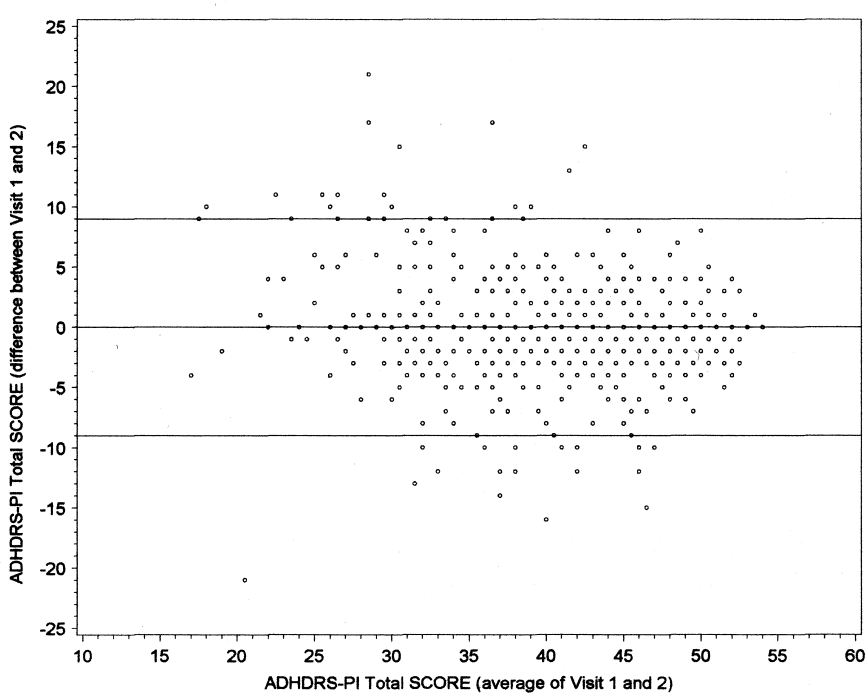

Figure 1. ADHDRS-PI: difference vs. average of total scores measured at Visits 1 and 2 with 95% limits of agreement.

change-from-baseline-to-endpoint measurements are presented in Table 4. Correlations between the ADHDRS-PI scores and other parent (CPRS) and clinician (CGI-ADHD-S) measures of ADHD symptom severity for both baseline and change score were moderate to high. Correlations with teacher-rated measures were low to moderate. All correlations were statistically different from 0.

Pearson's correlation coefficients between ADHDRS-PI scores and scores from other scales thought to measure a different construct at baseline, as well as the correlations between scales thought to measure the same set of symptoms, are presented in Table 5. As expected, the ADHDRS-PI inattentive subscale had a much lower correlation with other hyperactivity subscales than with assessments of other cognitive subscales. The corresponding trend was observed for the ADHDRS-PI hyperactive/impulsive subscale.

Discriminant validity

The ADHDRS-PI total score and subscale scores for patients with each ADHD subtype were summarized at baseline and reported in Figure 2. A statistically significant difference ($P < 0.001$) was noted between inattentive subtype patients and hyperactive/impul-

Table 4. Convergent validity of ADHDRS-PI based on score at baseline and change score from baseline to endpoint

Variable 1	Variable 2	Baseline			Change		
		N	Correlation[a]	(LCI[b], UCI[c])	*N*	Correlation[a]	(LCI[b], UCI[c])
ADHDRS-PI Total	CGI-ADHD-S	604	0.56	(0.50, 0.61)	603	0.77	(0.74, 0.80)
	CPRS ADHD Index	602	0.49	(0.42, 0.54)	566	0.71	(0.67, 0.75)
	CTRS ADHD Index	535	0.21	(0.13, 0.29)	457	0.18	(0.09, 0.27)
ADHDRS-PI	CPRS Cognitive	602	0.53	(0.47, 0.58)	566	0.67	(0.63, 0.72)
Inattentive	CTRS Cognitive	529	0.15	(0.07, 0.24)	452	0.14	(0.05, 0.23)
ADHDRS-PI	CPRS Hyperactive	603	0.73	(0.69, 0.76)	572	0.72	(0.68, 0.76)
Hyper/imp	CTRS Hyperactive	536	0.36	(0.28, 0.43)	461	0.21	(0.12, 0.29)

[a]Correlation is assessed by Pearson's correlation coefficient between variable 1 and variable 2.
[b]LCI – lower bound of 95% confidence interval based on Fisher's Z transformation.
[c]UCI – upper bound of 95% confidence interval based on Fisher's Z transformation.

Table 5. Correlations between the ADHDRS-PI subscales and corresponding CPRS and CTRS subscales at baseline

Variable 1	Variable 2			
ADHDRS-PI	CPRS		CTRS	
	Cognitive subscale correlation[a] (LCI[b], UCI[c]) *N* = 602	Hyperactive subscale correlation[a] (LCI[b], UCI[c]) *N* = 603	Cognitive subscale correlation[a] (LCI[b], UCI[c]) *N* = 529	Hyp/impulsive subscale correlation[a] (LCI[b], UCI[c]) *N*=536
Inattentive subscale	0.53 (0.47, 0.58)	0.15 (0.07, 0.23)	0.15 (0.07, 0.24)	0.02 (–0.07, 0.10)
Hyper/imp subscale	0.06 (–0.02, 0.14)	0.73 (0.69, 0.76)	–0.09 (–0.17, –0.01)	0.36 (0.28, 0.43)

[a]Correlation is assessed by Pearson's correlation coefficient between variable 1 and variable 2.
[b]LCI – lower bound of 95% confidence interval based on Fisher's Z transformation.
[c]UCI – upper bound of 95% confidence interval based on Fisher's Z transformation.

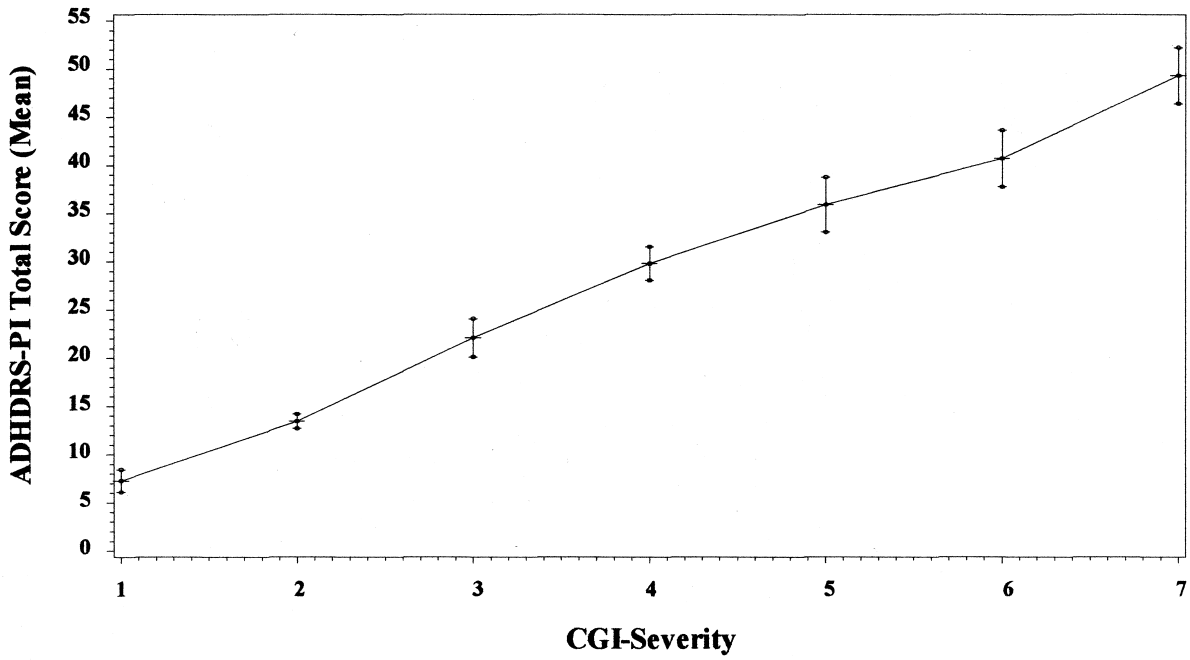

Figure 2. ADHDRS-PI baseline total and subscale scores by subtype.

sive subtype patients on both subscales of the ADHDRS-PI. In addition, statistically significant differences were noted between the inattentive-subtype patients and combined-subtype patients on the ADHDRS-PI hyperactive/impulsive subscale, and between the hyperactive/impulsive-subtype and combined-subtype patients on the ADHDRS-PI inattention subscale.

Figure 3 summarizes the mean ADHDRS-PI total scores by CGI-ADHD-S score at endpoint. Analyses of variance (ANOVA) on endpoint data indicate statistically significant differences in mean ADHDRS-PI total scores between each CGI-ADHD-S level (results not shown). In the total sample of patients, CGI-ADHD-S scores of 'minimally ill', 'mildly ill', 'moderately ill', and 'markedly ill' corresponded to mean ADHDRS-PI total t-score 52.3, 60.0, 68.6, and 78.1, respectively. In a population sample, a t-score of 50 represents the mean raw score for a child of a given age and gender, and each change of 10 points in t-score corresponding to 1 SD from the mean for each patient. In this study, normative data based on a sample of 2000 US children have been used to compute t-scores (transformations of raw scores based on normative data).

The Pearson correlation coefficients for total scores at baseline between ADHDRS-PI total score were –0.05 (P = 0.183) for CDRS, 0.016 (P = 0.709) for CDI, and 0.002 (P = 0.956) for MASC. Thus, correlations between the ADHDRS-PI total score and measurements of comorbid disease severity scores were very low and were not statistically significant.

Responsiveness

For patients who received at least one dose of atomoxetine, the mean baseline, mean change from baseline to endpoint, standard deviation in change scores, and the standardized response mean (SRM) for ADHDRS-PI, CGI-ADHD-S, CPRS, and CTRS are presented in Table 6. The SRM produced by the ADHDRS-PI scores was similar or numerically higher than the SRM using other parent and clinician measures of ADHD. All scales demonstrated a statistically significant change from baseline. SRM for ADHDRS-PI was also calculated for each ADHD subtype, and is included in Table 6. The scores were consistent across 3 subtypes, with the hyperactive/impulsive subtype having a slightly lower score, which might be due, in part, to the much smaller sample size. In addition, patients were divided into 3 groups according to their baseline ADHD RS total score, and the SRM for ADHDRS-PI was computed for each subgroup. The results are as follows: for Group 1 (ADHD RS total score (≤36), SRM =

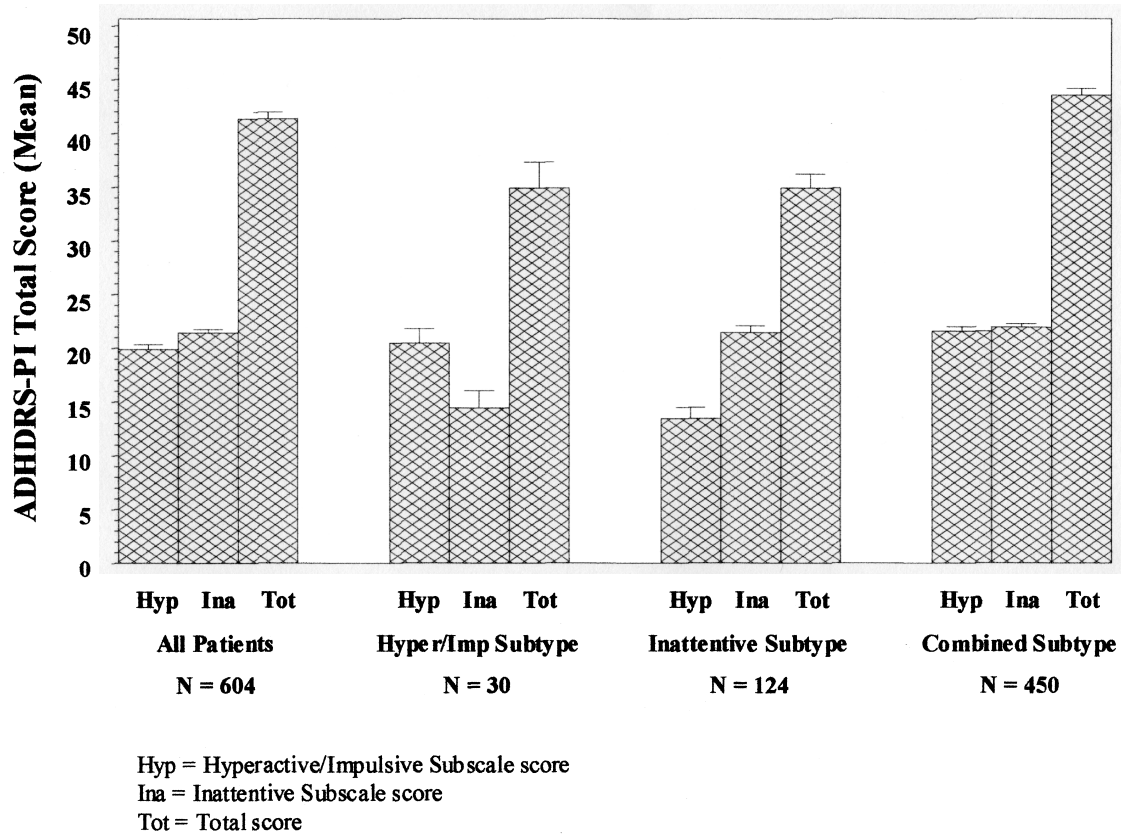

Figure 3. Mean ADHDRS-PI total scores by CGI-ADHD-severity score.

Table 6. Responsiveness of ADHDRS-PI based on change score from baseline to endpoint in the atomoxetine group

		Baseline			Change			
Population	Variable	N	Mean	SD	Mean	SD	P Value*	SRM**
All patients	ADHDRS-PI total	603	41.32	7.85	–23.37	12.68	<.001	1.84
Mixed subtype	ADHDRS-PI total	449	43.52	6.80	–24.57	12.90	<.001	1.90
Inatt. subtype	ADHDRS-PI total	124	34.89	7.33	–20.54	11.44	<.001	1.80
Hyper. subtype	ADHDRS-PI total	30	34.90	6.40	–17.13	10.64	<.001	1.61
All patients	CGI-ADHD-S	603	5.21	0.79	–2.65	1.46	<.001	1.82
All patients	CPRS ADHD index	572	28.37	5.70	–12.83	8.91	<.001	1.44
All patients	CTRS ADHD index	458	24.04	8.02	–5.98	7.70	<.001	0.78

*Within-treatment group P values are from Wilcoxon's Signed Rank Test.
**SRM (Standardized Response Mean) is computed as the mean change from baseline to endpoint divided by the standard deviation of change from baseline to endpoint scores.

1.67; for Group 2 (ADHD RS total score of 37 to 44), SRM = 1.96; and for Group 3 (ADHD RS total score (≤45), SRM = 2.21. It seems that SRM is affected by baseline severity; the higher the baseline ADHD RS total score, the greater the change from baseline to endpoint ($P < 0.001$).

Minimal clinically important differences

The within-treatment MCID for ADHD RS total score is 10.2 (or a 27% decrease from baseline). The between-treatment MCID is 6.6 (or a 19% decrease from baseline). If grouping the patients into approximately three equal groups based on their baseline ADHD RS total score (Group 1: (≤36, Group 2: 37–44, Group 3: (≤45), then the within-treatment MCID for the ADHD RS total score is 9.7 (or a 35% decrease from baseline) for Group 1, 9.6 (or a 24% decrease from baseline) for Group 2, and 11.4 (or a 23% decrease from baseline) for Group 3. The between-treatment MCID for the ADHD RS total score is 7.7 for Group 1, 7.6 for Group 2, and 5.2 for Group 3.

Ceiling and floor effects

The frequency of the highest and lowest possible scores is reported at both baseline and endpoint (Table 7). The percentage of patients achieving the highest and lowest possible scores was very small at both baseline and endpoint (less than 7%).

Validation of ADHDRS-PI by country

This was a multicountry study so the psychometric properties of ADHDRS-PI were also investigated for individual countries. Table 8 summarizes the key results.

Across all 14 countries, the range of Cronbach's alpha for the ADHDRS-PI total score at Visit 1 was 0.718 to 0.844. Intra-class correlation coefficients

Table 7. Frequency of best and worst possible scores for ADHDRS-PI at baseline and endpoint

	Baseline		Endpoint	
Frequency	Best score *n* (%)	Worst score *n* (%)	Best score *n* (%)	Worst score *n* (%)
ADHDRS-PI total	0 (0.0)	12 (1.86)	5 (0.78)	0 (0.0)
ADHDRS-PI inattentive	0 (0.0)	44 (6.82)	16 (2.48)	1 (0.16)
ADHDRS-PI hyper/imp	1 (0.16)	42 (6.51)	43 (6.67)	1 (0.16)

Table 8. Psychometric properties of ADHDRS-PI by country

Country	*N*	Internal Consistency (Cronbach's Alpha)	Test-retest reliability (ICC)	Responsiveness (SRM)
Australia	38	0.844	0.817	1.53
Belgium	36	0.815	0.890	1.31
Germany	26	0.718	0.899	1.73
Spain	36	0.765	0.865	2.42
France	35	0.736	0.760	1.82
UK	29	0.788	0.733	2.92
Hungary	77	0.741	0.768	2.68
Israel	34	0.815	0.871	1.15
Italy	31	0.718	0.804	2.37
Holland	31	0.798	0.731	1.33
Norway	24	0.731	0.946	1.74
Poland	62	0.739	0.868	2.03
Sweden	37	0.756	0.940	3.15
South Africa	68	0.815	0.800	2.02

between ADHDRS-PI total score for Visits 1 and 2 ranged from 0.731 to 0.946, respectively. The SRM produced for the ADHDRS-PI total scores had a minimum of 1.15 (Israel) and a maximum of 3.15 (Sweden). Positive correlations were found between ADHDRS-PI and other ADHD scales (CGI-ADHD-S and CPRS) for all 14 countries with the exception of Norway, where the correlation between ADHDRS-PI total score and CPRS ADHD Index subscale score was –0.08 at baseline. The number of patients with a hyperactive/impulsive subtype was too small to help assess discriminant validity within each country, a statistically significant difference ($P < 0.05$) was noted between inattentive-subtype patients and combined subtype patients on the ADHDRS-PI hyperactive/impulsive subscale. Correlations with the ADHDRS-PI inattentive subscale were much lower with CPRS Hyperactivity subscales than with assessments of CPRS cognitive subscales, and vice versa for the ADHDRS-PI hyperactive/impulsive subscale (results not shown here).

Discussion

In this study, the psychometric properties of the ADHD RS-PI, clinician administered and scored, were assessed in a group of over 600 patients from Europe (about 500 patients), Australia, Israel, and South Africa (106 patients) with a diagnosis of ADHD.

Generally, a Kappa statistic greater than 0.6 suggested good agreement, and a larger value indicated greater strength of agreement (Landis and Koch, 1977). There are no published guidelines that define an acceptable level of average squared deviation to assess inter-rater reliability. However, average squared deviation from a similar training tape for a US trial using ADHDRS-PI was 0.28 (Faries, 2001). These results suggested that the inter-rater reliability of ADHDRS-PI (Kappa = 0.63, average squared deviation = 0.30) is satisfactory for a multicentre clinical trial.

The results of exploratory factor analysis of the scale indicate that a two-factor solution would best represent the structure of this scale, which is also consistent with the DSM-IV two-dimensional diagnostic criteria.

To assess internal consistency, the literature suggests that 0.65 to 0.70 is an acceptable minimal standard (Nunnally, 1994; Perrin et al., 1997). The higher Cronbach's alpha, the greater the internal consistency. A very low value indicates that the rating scale is either too short or the items included in the scale have very little in common. Conversely, a very high value suggests some redundancy in the scale. Therefore, the internal consistency of the ADHDRS-PI scale (Cronbach's alpha = 0.795 at Visit 1 and 0.838 at Visit 2) is satisfied and is not sufficiently high to suggest redundancy in items high enough to suggest acceptable internal consistency. The modest item-to-total correlation also indicated that there were no items within the scale that showed either a non-acceptable correlation, or high colinearity. Note that Cronbach's alpha is a little higher at Visit 2 compared with Visit 1, which is consistent with the previous finding that the Cronbach's alpha value increases once raters have more experience with the scale. Also noted that this Cronbach's alpha is slightly lower than the target (0.90) suggested for use of the scale for individual rather than group comparisons (Perrin et al., 1997).

In this study, ICCs for ADHDRS-PI total and subscale scores range from 0.773 to 0.887, respectively, (see Table 2). Landis and Koch (1977) suggested that ICCs above 0.60 indicate satisfactory test-retest reliability and ICCs greater than 0.80 are excellent. The score of 0.773 to 0.887 indicated excellent test-retest reliability of ADHDRS-PI. Bland and Altman plot also showed a good agreement between scores obtained at two different time points, which suggested there is no potential systematic difference.

In general, correlations between the ADHDRS-PI and other measures of the same set of ADHD symptoms were high except for correlations with CTRS, especially for the ADHDRS-PI Inattentive subscale and CTRS cognitive subscale (see Table 3). At the same time, correlations between scales thought to measure different symptom groups were low. These results suggest adequate content validity for the scale. The low correlation between ADHDRS-PI and CTRS is consistent with other researches comparing parent and teacher scales (Faries, 2001). This may be due to multiple factors, including the fact that the scales are not completed at the same time (due to teacher schedules), there is limited contact or time to develop relationships between teacher and child, and the CTRS cognitive subscale assesses a slightly broader set of symptoms (academic performance) than just the ADHD symptom list. Previous studies also suggest that there is a low correlation between teachers' ratings and ratings made by trained classroom observers using the CTRS (Conger et al., 1983; Kazdin et al., 1983). This

may be in part due to the observation that teachers tend to put excessive emphasis on children's academic excellence (Lessing et al., 1974).

There were statistically significant differences ($P < 0.001$) among different ADHD-subtype patients (see Figure 1) indicating that the ADHDRS-PI discriminates between patients whose diagnosis suggested the presence of clinically significant hyperactive/impulsive or inattentive symptoms and those whose diagnosis did not. The statistically significant differences in ADHDRS-PI mean total scores between each CGI-Severity score suggest that ADHDRS-PI had the ability to discriminate between groups of patients at different severity levels (see Figure 2). These two results indicate the satisfactory discriminant validity of the ADHDRS-PI scale. Patients with ADHD showed large improvement in ADHDRS-PI total score over time (SRM ranged from 1.61 to 1.90 for patients with different ADHD subtype). There are no published guidelines that define an acceptable level of SRM. However, this result is comparable with CGI-ADHD-S (SRM = 1.82), which is a well-accepted clinician rating scale. This is also consistent with the result (SRM = 1.21) from a US trial with a similar design using the ADHDRS-PI (Faries, 2001). These results indicated acceptable responsiveness of the ADHDRS-PI scale. Note that the SRM score in this study is relatively high. We think the reason is that most patients entered the study with severe ADHD symptoms (mean ADHD RS total t-score 3 SD above norms), and thus had bigger opportunity to change. The higher SRM for patients with a higher baseline score supports our assumption.

For ADHDRS-PI total score, the percentages of patients given the best and worst possible scores were very small (see Table 7), thus suggesting that ceiling and floor effects are not an issue for this scale in this patient population.

The patient characteristics and baseline ADHD severity scores in this study were very similar to those enrolled in a similar US study used to assess the validity of ADHDRS-PI (Faries, 2001). Table 1 summarizes the patient baseline characteristics for both studies. Moreover, the psychometric properties of the scale for both patient populations were similar. This similarity allows us to address the interpretation of data from US studies relative to other countries. The consistency of the results of this study with the results of the atomoxetine study conducted in the US suggests that ADHDRS-PI can be successfully used to identify patients with ADHD symptoms, and to detect the change of symptom severity over time under treatment in both the US and outside the US.

The ADHDRS-PI demonstrated acceptable internal consistency across all 14 countries (minimum Cronbach's alpha was 0.718, see Table 8). ADHDRS-PI was also found to have acceptable levels of test-retest reliability for all 14 countries as ICCs ranged from 0.731 to 0.946. Note correlations of at least 0.60 are considered satisfactory while those greater than 0.80 are considered excellent (Landis and Koch, 1977). As with findings from the overall population, acceptable levels of convergent/divergent validity were found between ADHDRS-PI and CPRS for all 14 countries. Acceptable discriminant validity was observed within each country. The SRM produced for the ADHDRS-PI total scores by countries had the minimum as 1.15, which indicated acceptable responsiveness of ADHDRS-PI for all countries. Consistent results of the psychometric properties across patient groups from multiple countries in this study indicate this rating scale is reasonable for multinational practice.

There are several limitations to this study. It has been noted that culture may play a role in ADHD assessment. Though Magnusson et al. (1999) showed that the factor structures of this rating scale were highly similar across cultures and the norm scores were similar to those found in American studies when rated by parents, that study used an Icelandic version of the scale and only applied it to Icelandic schoolchildren. In this study, the English version of the rating scale was utilized in countries where English was not the native language and the US norms were applied to populations outside the US. Thus, the ability of the clinicians to translate to non-English-speaking patients was an additional source of variability and results may not extend to other or future translations. Nonetheless, Buitelaar et al. (2004) reported that while there were several differences between the international and North American study populations, the two groups were very similar in most respects.

We also realized that because of the small sample size in each country (604 participants spread over 14 countries), it is difficult to interpret the validity of this rating scale for each country.

In this study, children without ADHD symptoms were not recruited for ethical reasons. Therefore, a

comparison between patients with ADHD and controls was not possible.

Although discriminant validity could be assessed by comparing patients with a diagnosis of ADHD inattentive subtype and a diagnosis of ADHD combined subtype (defined elsewhere in text), the inference was somewhat limited. Additionally, there was neither a placebo-control arm nor an active comparator (methylphenidate) in this analysis. Therefore, we could not assess responsiveness of the scale using effect sizes based on treatment differences from placebo, nor compare the standardized response mean from atomoxetine with another proven efficacious compound.

Another limitation is that more than 50% of patients had previous drug therapy for ADHD, and we speculate that the parents of those patients with previous medication might have significantly more knowledge about ADHD than parents of children without previous medication. Clinician-scored ADHDRS-PI is based on interviews with parents, so there is potential bias toward either higher or lower scores for patients with previous experience compared with patients who were treatment naïve. This aspect is not covered in this paper and might be worth further research.

In conclusion, our results support the validity and reliability of the ADHD RS as a clinician-administered and clinician-scored tool for assessing the severity of ADHD symptoms in children and adolescents under a research setting in Europe. The ADHD RS as a clinician-rated tool could be used for screening patients for baseline symptom severity to determine whether they meet a threshold condition for study participation and for following the course of symptom change over a clinical trial.

Acknowledgement

This research was funded by Eli Lilly and Company.

References

Allen AJ, Spencer TJ, Heiligenstein JH, Faries DE, Kelsey DK, Laws HF, Wernicke J, Kendrick KL, Michelson D. Safety and efficacy of atomoxetine for ADHD in two double-blind, placebo-controlled trials. Biological Psychiatry, Society of Biological Psychiatry 56th Annual Convention and Scientific Program, 3–5 May 2001; New Orleans LA. Dallas, TX: Elsevier.

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR. Washington DC: American Psychiatric Press, 2000.

Anderson TW. An Introduction to Multivariate Statistical Analysis. 2 edn. Beijing People's Republic of.China: John Wiley & Sons, 1990, pp. 120–5.

Barkley RA. Attention Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment. New York: Guilford, 1990.

Baumgaertel A, Wolraich ML, Dietrich M. Comparison of diagnostic criteria for attention deficit disorders in a German elementary school sample. J Am Acad Child Adolesc Psychiatry 1995; 34: 629–38.

Biederman J, Faraone SV, Monuteaux MC, Grossbard JR. How informative are parent reports of attention-deficit/hyperactivity disorder symptoms for assessing outcome in clinical trials of long-acting treatments? A pooled analysis of parents' and teachers' reports. Pediatrics 2004; 113: 1667–71.

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; I: 307–10.

Brown, RT, Freeman WS, Perrin JM, Stein MT, Amler RW, Feldman HM, Pierce K, Buitelaar JK, Danckaerts M, Gillberg C, Zuddas A, Becker K, Bouvard M, Fagan J, Gadoros J, Harpin V, Hazell P, Johnsson M, Lerman-Sagie T, Soutullo C, Wolanczyk T, Zeiner P, Fouche DS, Krikke-Workel J, Zhang S, Michelson. A prospective, multicenter, open-label assessment of atomoxetine in non-North American children and adolescents with ADHD. Eur Child Adolesc Psychiatry 2004; 13(4): 249–57.

Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin. 1959; 56: 81–105.

Channon E, Butler A. Comparing investigators' use of rating scales such as PANSS in muti-investigator studies of schizophrenia. Poster presentation at the Eleventh European College of Neuropsychopharmacology (ECNP) Congress, 1998.

Cohen M, Becker MG, Campbell R. Relationships among four methods of assessment of children with attention deficit-hyperactivity disorder. J School Psychol 1990; 28: 189–202.

Conger AJ, Conger JC, Wallander J, Wark D, Dygdon J. A generalizability study of the Conners' Teacher Rating Scale-Revised. Educ Psychol Meas 1983; 43: 1019–31.

Conners CK. Conners' Rating Scales: Revised Technical Manual. North Towanda (New York): Multi-Health Systems, 1997.

Cronbach LF. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16: 297–334.

Deyo RA, Diehr P, Patrick KL. Reproducibility and responsiveness of health status measures: Statistics and stratifies for evaluation. Control Clin Trials 1991; 12: 142S–158S.

DuPaul GJ, Power TJ, Anastopoulos AD, Reid R. ADHD Rating Scale IV: checklists, norms, and clinical interpretation. New York: Guilford, 1998.

Faraone SV, Biederman J. Neurobiology of attention-deficit hyperactivity disorder. Biol Psychiatry 1998; 10: 951–8.

Faries DE, Yalcin I, Harder D, Heiligenstein JH. Validation of the ADHD Rating Scale as a clinician administered and scored instrument. J Atten Disord 2001; 5(2): 107–15.

Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971; 76(5): 378–82.

Graetz BW, Sawyer MG, Hazell PL, Arney F, Baghurst P. Validity of DSM-IV ADHD subtypes in a nationally representative sample of Australian children and adolescents. J Am Acad Child Adolesc Psychiatry 2001; 40: 1410–17.

Guy W. ECDEU assessment manual for psychopharmacology, revised. Bethesda (MD): US Department of Health, Education, and Welfare, 1976.

Health Council of the Netherlands. Diagnosis and treatment of ADHD. Publication no. 2000/24. The Hague: Health Council of the Netherlands, 2000.

Herrmann C. International experiences with the Hospital Anxiety and Depression Scale - a review of validation data and clinical results. J Psychosom Res 1997; 42: 17–41.

Kaufman J, Birmaher B, Brent D, Rao U, Ryan N. Kiddie-Sads-Present and Lifetime Version (K-SADS-PL). Version 1.0. Pittsburgh: Department of Psychiatry, University of Pittsburgh School of Medicine, 1996.

Kazdin AE, Esveldt-Dawson K, Loar LL. Correspondence of teacher ratings and direct observations of classroom behaviour of psychiatric inpatient children. J Abnorm Psychol 1983; 10; 483–95.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159–74.

Lessing E, Oberlander MI, Barbera L. Convergent validity of the IPAT Children's Personality Questionnaire and teachers' ratings of the adjustment of elementary school children. Soc Behav Pers 1974; 2: 222–9.

Magnusson P, Smart J, Gretarsdottir H, Prandardottir H. Attention-Deficit/Hyperactivitity symptoms in Icelandic schoolchildren: assessment with the Attention Deficit/Hyperactivity Rating Scale-IV. Scand J Psychol 1999; 40: 301–6.

Livingston R. Cultural issues in Diagnosis and treatment of ADHD. J Am Acad Child Adolesc Psychiatry 1999; 38(12): 1591–4.

Michelson D, Faries DE, Wernicke J, Kelsey DK, Kendrick KL, Sallee FR, Spencer T, and the Atomoxetine ADHD Study Group. Atomoxetine in the treatment of children and adolescents with ADHD: A randomized, placebo-controlled dose-response study. Pediatrics 2001; 108 (5): 1–9.

Michelson D, Buitelaar JK, Danckaerts MJ, Gillberg C, Spencer T, Zuddas A, Faries D, Zhang S, Biederman J. Relapse prevention in Patients with Attention-Deficit/Hyperactivity disorder treated with atomoxetine: a randomized, double-blind, placebo-controlled study. J Am Acad Child Adolesc Psychiatry 2004; 43(7): 896–904.

Mueller CW, Mann EM, Thanapum S, Humris E, Ikeda Y, Takahashi A, Tao KT, Li BL. Teachers' ratings of disruptive behaviour in five countries. J Clin Child Psychol 1995; 24(4): 434–42.

National Institute of Clinical Excellence. Guidance on the Use of Methylphenidate (Ritalin, Equasym) for Attention Deficit/Hyperactivity Disorder (ADHD) in Childhood. Technology Appraisal Guidance no. 13. London: NICE, 2000.

Nunnally, JC. Psychometric Theory. 3 edn. New York: McGraw-Hill, 1994.

Perrin ES, Aaronson NK, Alonso J, Burnam A, Lohr K, Patrick KL. Instrument Review Criteria. Medical Outcomes Trust 1997: 1–5.

Pliszka SR, McCracken JT, Maas JW. Catecholamines in attention-deficit hyperactivity disorder: current perspectives. J Am Acad Child Adolesc Psychiatry 1996; 35(3): 264–72.

Prendergast M, Taylor E, Rapoport JL, Bartko J, Donnelly M, Zametkin A, Ahearn MB, Dunn G, Wieselberg HM. The diagnosis of childhood hyperactivity. A US-UK cross-national study of DSM-III and ICD-9. J Child Psychol Psychiatry 1988; 29(3): 289–300.

Reid, R. Assessment of ADHD with culturally different groups: the use of behaviour rating scales. School Psychology Review 1995; 24: 537–60.

Reid R, DuPaul GJ, Power TJ, Anastopoulos AD, Rogers-Adkinson D, Noll M, Riccio C. Assessing culturally different students for attention deficit hyperactivity disorder using behaviour rating scales. J Abnorm Child Psychol 1998; 26(3): 187–98.

Spencer F, Biederman J, Heiligenstein J, Wilens T, Faries D, Prince J. An open-label, dose ranging study of tomoxetine in children with attention deficit hyperactivity disorder. J Child Adolesc Psychopharmacol 2001; 11(3): 251–65.

Stratford PW, Binkley JM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. Phys Ther 1996; 76(10): 109–23.

Stucki G, Michel BA. How to measure improvement: rules and fallacies. Rheumatol Eur Supp 1995; 2: 107–11.

Correspondence: Shuyu Zhang, Lilly Research Laboratories, Eli Lilly and Company, Drop Code 6161, Indianapolis, IN 46285.

Telephone (+1) 317-2763455.
Fax (+1) 317-4336590.
Email: shuyu_zhang@lilly.com.