Psychological Medicine, 2005, **35**, 245–256. © 2004 Cambridge University Press
DOI: 10.1017/S0033291704002892 Printed in the United Kingdom

The World Health Organization adult ADHD self-report scale (ASRS): a short screening scale for use in the general population

RONALD C. KESSLER*, LENARD ADLER, MINNIE AMES, OLGA DEMLER, STEVE FARAONE, EVA HIRIPI, MARY J. HOWES, ROBERT JIN, KRISTINA SECNIK, THOMAS SPENCER, T. BEDIRHAN USTUN AND ELLEN E. WALTERS

Department of Health Care Policy, Harvard Medical School; Departments of Psychiatry and Neurology, New York University School of Medicine; Department of Psychiatry, Massachusetts General Hospital; Eli Lilly and Company, Global Health Outcomes; Global Burden of Disease Unit, World Health Organization

ABSTRACT

Background. A self-report screening scale of adult attention-deficit/hyperactivity disorder (ADHD), the World Health Organization (WHO) Adult ADHD Self-Report Scale (ASRS) was developed in conjunction with revision of the WHO Composite International Diagnostic Interview (CIDI). The current report presents data on concordance of the ASRS and of a short-form ASRS screener with blind clinical diagnoses in a community sample.

Method. The ASRS includes 18 questions about frequency of recent DSM-IV Criterion A symptoms of adult ADHD. The ASRS screener consists of six out of these 18 questions that were selected based on stepwise logistic regression to optimize concordance with the clinical classification. ASRS responses were compared to blind clinical ratings of DSM-IV adult ADHD in a sample of 154 respondents who previously participated in the US National Comorbidity Survey Replication (NCS-R), oversampling those who reported childhood ADHD and adult persistence.

Results. Each ASRS symptom measure was significantly related to the comparable clinical symptom rating, but varied substantially in concordance (Cohen's κ in the range 0·16–0·81). Optimal scoring to predict clinical syndrome classifications was to sum unweighted dichotomous responses across all 18 ASRS questions. However, because of the wide variation in symptom-level concordance, the unweighted six-question ASRS screener outperformed the unweighted 18-question ASRS in sensitivity (68·7% *v.* 56·3%), specificity (99·5% *v.* 98·3%), total classification accuracy (97·9% *v.* 96·2%), and κ (0·76 *v.* 0·58).

Conclusions. Clinical calibration in larger samples might show that a weighted version of the 18-question ASRS outperforms the six-question ASRS screener. Until that time, however, the unweighted screener should be preferred to the full ASRS, both in community surveys and in clinical outreach and case-finding initiatives.

INTRODUCTION

Although it has long been known that attention-deficit/hyperactivity disorder (ADHD) is one of the most common psychiatric disorders among children (Shekim *et al.* 1985; Bird *et al.* 1988) and that ADHD often persists into adulthood (Menkes *et al.* 1967; Mannuzza *et al.* 1993), the fact that adult ADHD is a commonly occurring and seriously impairing disorder has only recently become the focus of attention (Wender *et al.* 2001; Pary *et al.* 2002; Wilens *et al.* 2002). One commentator has gone so far as to suggest

* Address for correspondence: Dr R. C. Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA.
(Email: kessler@hcp.med.harvard.edu)

that ADHD is probably 'the most common chronic undiagnosed psychiatric disorder in adults' (Wender, 1998). However, no large-scale epidemiological data exist to evaluate this claim, as none of the many adult community psychiatric epidemiological surveys carried out over the past two decades with either the Diagnostic Interview Schedule (DIS; Robins *et al.* 1981) or the Composite International Diagnostic Interview (CIDI; Robins *et al.* 1988) included an assessment of adult ADHD.

Several attempts have been made to estimate the general population prevalence of adult ADHD by extrapolation from childhood prevalence estimates in conjunction with adult persistence estimates (Weiss *et al.* 1985; Mannuzza *et al.* 1998; Biederman *et al.* 2000; Barkley *et al.* 2002) or by direct estimation from small samples of adults (Murphy & Barkley, 1996) or college students (Heiligenstein *et al.* 1998), yielding prevalence estimates in the range 1–6%. However, these estimates are all based on convenience samples. In an effort to obtain more representative estimates, adult ADHD is included in the new World Health Organization (WHO) World Mental Health (WMH) Initiative surveys (Kessler & Ustun, 2000), a series of general population psychiatric epidemiological surveys currently underway in 28 different countries in all regions of the world with a combined sample size of more than 200 000 respondents.

Both retrospective assessments of childhood ADHD and a screen for adult ADHD were developed for use in the expanded version of CIDI that forms the core of the WMH surveys. The retrospective assessment was part of a series of four sections devoted to adult recall of childhood disorders (oppositional-defiant disorder, conduct disorder, and separation anxiety disorder in addition to ADHD) that were based on the comparable sections in the DIS (Robins *et al.* 1995). There was no precedent, though, for assessing adult ADHD in previous DIS-CIDI surveys. Although other self-report measures of adult ADHD exist (Barkley, 1995; Brown, 1996; Conners *et al.* 1998; Mehringer *et al.* 2002; West *et al.* 2003), a review of these measures showed that they either fail to include all 18 DSM-IV Criterion A symptoms or assess some of these symptoms with questions that were judged to be suboptimal by an Advisory Group of clinical experts in adult ADHD assembled by the WHO to consult on this aspect of the WMH survey assessment. Based on this evaluation, a decision was made to develop a new self-report measure of adult ADHD for the WMH surveys. The present report describes this measure: the WHO Adult ADHD Self-Report Scale (ASRS) Version 1.1. A short ASRS screener, which turned out to outperform the full ASRS in the clinical calibration study reported here, was also developed based on stepwise logistic regression.

METHOD

Participants

The clinical calibration of the ASRS was carried out by re-interviewing a quota subsample of 154 respondents from the US National Comorbidity Survey Replication (NCS-R; Kessler *et al.* 2003), a nationally representative face-to-face household survey of 9083 respondents aged 18 and older in the co-terminous United States who were interviewed by trained lay interviewers between February 2001 and December 2002. NCS-R respondents were selected from a stratified, multi-stage, clustered area probability sample of the USA non-institutionalized civilian population, with over 1000 segments (i.e. block-equivalents) in 170 counties in 34 states. The response rate among primary respondents was 70·9%. All respondents were administered the WMH version of the CIDI in Part I of the interview, while a probability subsample of 5692 respondents also received a Part II interview that included assessments of risk factors and other disorders. The Part II sample consisted of all respondents who screened positive for any Part I WMH-CIDI disorder plus a probability subsample of 25% other Part I respondents. Less than 1% of the Part I respondents who were selected into the Part II sample failed to complete Part II. This conditional response rate was unrelated to the Part II sampling stratum. A more detailed description of the NCS-R sample design and field procedures is presented elsewhere (Kessler *et al.* in press).

Based on a concern that older adults would have difficulty responding to retrospective questions about childhood, the assessment of ADHD was limited to respondents in the age range 18–44 years. Respondents who reported symptoms that were classified as having met

criteria for ADHD in childhood were asked a single follow-up question about whether they continued to have any current problems with attention or hyperactivity-impulsivity. The respondents who received the ADHD assessment were divided into four sampling strata for the adult ADHD clinical calibration sample: those who denied any childhood symptoms of ADHD, those who reported at least some symptoms but were classified as not meeting full criteria, those who were classified as meeting criteria but denied having current adult symptoms, and those who were classified as meeting criteria who reported having current adult symptoms.

An attempt was made to contact by telephone and re-interview 30 respondents in each of the first three strata and 60 in the fourth stratum. The final quota sample of 154 was slightly larger than the sum of these targets because a higher than expected proportion of pre-designated respondents kept their appointments to be interviewed. Before beginning the interviews, respondents were told that the purpose of the interviews was to test the accuracy and expand the measures in the original survey, that participation was completely voluntary and confidential, and that we would send respondents a check for $50 as a token of our appreciation for their participation in this phase of the research. Respondents were then allowed to ask any additional questions before obtaining verbal informed consent and beginning the interview. Interviews were tape-recorded with the permission of respondents. The Harvard Medical School Human Subjects Committee approved these recruitment and consent procedures.

Two levels of weighting were required to make the 154 respondents representative of the total NCS-R sample. First, the sample was weighted using the NCS-R Part II composite weight to adjust for differential probabilities of selection into the overall sample within households; differences in intensity of recruitment effort among hard-to-recruit cases; differential non-response across sample segments based on aggregated Census Block Group data; discrepancies between the sample and the Census population distribution on the cross-classification of various sociodemographic variables; and differential probabilities of selection into the Part II sample. The development of these weights is discussed in detail elsewhere (Kessler *et al.* 2003). Second, the sample was weighted to adjust for the oversampling of Part II respondents who reported childhood ADHD and adult persistence of ADHD symptoms. It has previously been shown that the weighted NCS-R Part II sample distribution closely matches the Census population on a variety of geographic and demographic variables (Kessler *et al.* 2003). The weighted ADHD clinical calibration sample distribution was found roughly to approximate the Part II NCS-R distribution on these same sociodemographic variables. (Appendix tables that include details of these comparisons along with other results that are too detailed to be reported in this paper are available at www.hcp.med.harvard.edu/ncs).

The screening scale item pool

Two board-certified psychiatrists (L.A., T.S.) and the WHO advisory group of clinical experts in adult ADHD (see Acknowledgments) generated an initial pool of fully structured questions about the symptoms of ADHD as they are typically expressed among patients with adult ADHD and mapped these onto each of the 18 DSM-IV Criterion A symptoms. The survey methodology collaborators (R.C.K., T.B.U.) then modified these questions to remove double-barreled descriptions, reduce ambiguities in meaning, and create a consistent temporal focus. The clinical collaborators then made modifications to improve face validity. One question was selected from this final item pool for each of the 18 DSM-IV Criterion A symptoms of ADHD based on face validity. Eleven more questions were selected to span the range of symptoms not in DSM-IV that were thought by the clinical experts to be common expressions of adult ADHD. Each question asked how often a symptom occurred over the past 6 months on a 0–4 scale with responses of never (0), rarely (1), sometimes (2), often (3), and very often (4). The focus of this report is on the 18 questions designed to operationalize the DSM Criterion A symptoms (Table 1).

The clinical interview

The clinical interview used in the calibration study had three parts. The first part was the semi-structured clinical ADHD Rating Scale (ADHD-RS; DuPaul *et al.* 1998), a state-of-the-art retrospective assessment of childhood

Table 1. *The WMH-CIDI Adult ADHD Self-Report Scale (ASRS) Questions*

I. Inattention		
	1.	How often do you make careless mistakes when you have to work on a boring or difficult project?
	2.	How often do you have difficulty keeping your attention when you are doing boring or repetitive work?
	3.*	How often do you have difficulty concentrating on what people say to you, even when they are speaking to you directly?
	4.*†	How often do you have trouble wrapping up the fine details of a project, once the challenging parts have been done?
	5.*†	How often do you have difficulty getting things in order when you have to do a task that requires organization?
	6.†	When you have a task that requires a lot of thought, how often do you avoid or delay getting started?
	7.	How often do you misplace or have difficulty finding things at home or at work?
	8.	How often are you distracted by activity or noise around you?
	9.*†	How often do you have problems remembering appointments or obligations?
II. Hyperactivity-Impulsivity		
	1.†	How often do you fidget or squirm with your hands or your feet when you have to sit down for a long time?
	2.*	How often do you leave your seat in meetings or other situations in which you are expected to remain seated?
	3.	How often do you feel restless or fidgety?
	4.	How often do you have difficulty unwinding and relaxing when you have time to yourself?
	5.†	How often do you feel overly active and compelled to do things, like you were driven by a motor?
	6.	How often do you find yourself talking too much when you are in a social situation?
	7.*	When you're in a conversation, how often do you find yourself finishing the sentences of the people that you are talking to, before they can finish them themselves?
	8.	How often do you have difficulty waiting your turn in situations when turn-taking is required?
	9.*	How often do you interrupt others when they are busy?

Response options are: never, rarely, sometimes, often, and very often. Patients were asked to answer the questions using a 6-month recall period.

* Clinically significant symptom levels were defined for these seven questions as responses of sometimes, often, and very often. For all remaining 11 questions, often and very often were the clinically significant symptom levels. See text for the rationale for this difference.

† The six-question ASRS screener.

ADHD designed for administration to adults. The second part was the semi-structured clinical interview for recent (past 6 months) DSM-IV adult ADHD that is used in most clinical trials of this disorder (Spencer *et al.* 1995, 1998, 2001; Michelson *et al.* 2003). The third part was the self-report battery, which was administered at the end in order not to bias interviewers when they made their clinical symptom ratings. Four experienced clinical interviewers (Ph.D. clinical psychologists with between 5 and 20 years of clinical experience) carried out these interviews. Each interviewer received 40 hours of training from two board certified psychiatrists who specialize in research on adult ADHD (L.A., T.S.) and successfully completed five practice interviews in which their symptom ratings matched those of the trainers before they began production interviewing.

Clinical supervisor (M.H.) reviewed tape recordings of clinical interviews. Discrepancies between interviewer and supervisor ratings were referred to the board certified psychiatrists (L.A., T.S.) for discussion and resolution and, as needed, additional contacts with respondents. The supervisor and clinical collaborators also held weekly group interviewer calibration meetings, while the supervisor held weekly one-on-one feedback meetings with each interviewer separately. Consistent with DSM-IV criteria, a clinical diagnosis of adult ADHD required that a respondent have at least six symptoms of either inattention or hyperactivity-impulsivity during the 6 months before the interview (DSM-IV Criterion A), at least two Criterion A symptoms of ADHD before age seven (Criterion B), some impairment in at least two areas of living during the past 6 months (Criterion C), and clinically significant impairment in at least one area of living over the same time period (Criterion D). No attempt was made to operationalize the DSM-IV diagnostic hierarchy rules for ADHD (Criterion E).

Statistical methods

Symptom-level concordance between ASRS self-reports and blind clinician ratings of 6-month prevalence was evaluated by dichotomizing the 0–4 ASRS response scale separately for each question to maximize overall classification accuracy with dichotomous clinical symptom ratings. A wide range of possible simple scoring methods that used all 18 questions was then investigated. Diagnostic efficiency statistics were calculated for each of these to select the best method, including sensitivity (the percent of respondents with the clinician-rated syndrome classified as having

the syndrome by the screening scale); specificity (the percent of respondents without the clinician-rated syndrome classified as not having the syndrome by the screening scale); total classification accuracy (the percent of all respondents consistently classified by the ASRS and clinician ratings); the odds ratio (OR) of the 2×2 table between yes–no syndrome classifications based on the ASRS and the clinical ratings; Cohen's κ (a measure of concordance that adjusts for chance agreement); and the area under the receiver operator characteristic curve (AUC; the probability that a randomly selected clinical case would score higher on the ASRS than a randomly selected non-case).

In order to investigate whether most of the precision of the full 18-question ASRS could be captured with fewer items, stepwise logistic regression analysis was used to select the best subset of ASRS questions to create a short-form screener. Only six questions were found to enter this stepwise analysis significantly. However, all-possible subsets logistic regression showed that a number of different six-question subsets of the 18 questions were roughly equivalent in reproducing clinical diagnoses. As a result, the psychometric analyses described in the last paragraph were repeated for all these alternative six-question short-form scales. Inspection of test statistics was used to select a final optimal short-form scale.

The conventional screening approach creates a dichotomy to differentiate predicted cases and non-cases. However, this dichotomization often discards potentially useful information that would be retained in a polychotomous screening scale, such as the distinction between a nearly definite case and a probable case. As a result, the use of polychotomous screening scales is becoming increasingly popular in evidence-based medicine (Peirce & Cornell, 1993). In order to investigate whether this might be useful for the ASRS, polychotomous versions of both the optimal 18-question and six-question scoring methods were created.

As the sample design features clustering and weighting of cases, design-based methods were used to calculate standard errors and confidence intervals. The jackknife pseudo-replication method (Wolter, 1985) implemented in an SAS 8 macro (SAS Institute, 1999) was used to make these calculations.

RESULTS

Symptom-level concordance

In order to assess symptom-level concordance, the 0–4 ASRS response scale was collapsed into a dichotomy that mapped onto the dichotomous clinical ratings. The conventional way to do this is to select the same dichotomy for all questions, with responses of often or very often usually considered above the clinical threshold and other responses (never, rarely, and sometimes) below the threshold (e.g. O'Donnell *et al*. 2001). However, this approach can lead to error either when the self-report questions vary in severity or when respondents are more reluctant to admit the frequent occurrence of some symptoms than others, in which case it is preferable to allow between-question variation in thresholds (Kessler *et al*. 2002). Based on this thinking, the 2×5 cross-classification of each 0–4 ASRS response with the dichotomous clinical symptom rating was inspected and the ASRS response scale was dichotomized in such a way that the numbers of false positives (ASRS positives who were rated as asymptomatic in the clinical interview) and false negatives (ASRS negatives who were rated as having the symptom in the clinical interview) in the weighted data were as equal as possible. This rule resulted in clinically significant symptom levels being defined as those that occurred often or very often for 11 screening questions and as sometimes, often, or very often for the other seven screening questions.

Cohen's κ was used to assess concordance between dichotomized ASRS symptom responses and clinical symptom ratings. Applying commonly used standards for assessing strength of κ coefficients (Landis & Koch, 1977), concordance was slight (less than 0·2) for two questions, fair (0·2–0·4) for seven, moderate (0·4–0·6) for six, and substantial (0·6–0·8) for the remaining three. Eleven of the 18 questions were found to be unbiased in the sense that the number of false positives did not differ significantly from the number of false negatives at the 0·05 level of significance using two-sided tests. The proportion of biased questions was comparable for inattention and hyperactivity-impulsivity. Three of the seven biased questions were biased downward and the other four were biased upward. Detailed results of these

Table 2. *Concordance of optimally dichotomized[a] versions of the 18-question ASRS with blind ADHD-RS clinical syndrome classifications using a variety of scoring methods*

	Sensitivity[c]		Specificity[c]		TCA[c]			McNemar test[d]	Cohen's kappa[c]		Odds ratio[c]		
Optimal cut-point[b]	%	(S.E.)	%	(S.E.)	%	(S.E.)	FP-FN[c]	χ^2_1	κ	(S.E.)	OR	(95% CI)	AUC[c]
(1) 0–5 *v.* 6–9	51·8	(9·0)	97·2	(2·0)	94·9	(2·1)	0·2	0·0	0·48	(0·16)	37·1*	(7·3–188·4)	0·74
(2) 0–6 *v.* 7–18	34·0	(8·4)	96·9	(2·1)	93·7	(2·2)	−0·4	0·0	0·32	(0·17)	15·9*	(3·4–73·8)	0·65
(3) 0–20 *v.* 21–36	60·2	(8·9)	96·3	(2·1)	94·5	(2·1)	1·5	0·6	0·50	(0·15)	39·5*	(9·9–157·0)	0·78
(4) 0–8 *v.* 9–18	56·3	(8·8)	98·3	(0·7)	96·2	(0·9)	−0·6	0·1	0·58	(0·16)	73·4*	(25·9–208·6)	0·77
(5) 0–7 *v.* 8–36	55·9	(9·1)	96·5	(2·1)	94·5	(2·1)	1·1	0·3	0·48	(0·15)	35·4*	(8·5–148·0)	0·76
(6) 0–36 *v.* 37–72	57·2	(9·1)	96·5	(2·1)	94·5	(2·1)	1·2	0·4	0·48	(0·15)	36·5*	(8·9–150·3)	0·77

[a] Optimality was defined as minimizing the difference between the weighted number of false positive (FP) and false negative (FN) responses.
[b] The first three scoring methods calculated separate scores for the inattention and the hyperactivity-impulsivity domains and assigned the higher of these two scores. The first method counted the number of positive symptom screens (defined in Table 3) in each domain. The second method assigned greater weight to responses of 'very often' than to other positive responses and then counted the weighted number of positive symptom screens in each domain, with a range of relative weights up to a maximum of 2 : 1. Only results of the 2 : 1 version are reported here. The third method included the full 0–4 range of responses to the ASRS questions and summed these responses separately for the inattention and hyperactivity-impulsivity domains. The fourth through sixth scoring methods were similar to the first three except that they were based on a summation across all 18 questions rather than on the higher score in the separate inattention and hyperactivity-impulsivity domains.
[c] See the text for definitions of the test statistics.
[d] The McNemar test evaluates the significance of the difference between FP and FN. None of these tests was significant at the 0·05 level.
TCA, total classification accuracy; AUC, area under the receiver operator characteristic curve.
* Significant at the 0·05 level.

comparisons are included in the appendix materials posted at the website mentioned earlier in the paper.

Optimal scoring of the ASRS

The most direct mapping of ASRS responses onto DSM-IV Criterion A is to classify respondents as cases if they have positive responses either to six or more inattention questions or to six or more hyperactivity-impulsivity questions. However, as concordance between the ASRS questions and clinical symptom ratings is far from perfect, application of the six-of-nine rule might not be optimal in predicting clinical syndrome classifications. A total of six dimensional scoring methods were consequently applied to the data and concordance with the clinical syndrome classifications assessed for all dichotomizations of each of these six. The first three of the six scoring methods were based on separate scores for the inattention and the hyperactivity-impulsivity domains, with each respondent receiving the higher of these two scores.

The first method counted the number of positive symptom screens in each domain (one dichotomization of which is the DSM-IV six-of-nine categorization).

The second method assigned greater weight to responses of 'very often' than to other positive responses and then counted the weighted number of positive symptom screens in each domain. A range of relative weights was considered here up to a maximum of 2 : 1.

The third method included the full 0–4 range of responses to the screening questions and summed these responses in each domain.

The fourth through sixth scoring methods were similar to the first three except that they were based on a summation across all 18 questions rather than on the higher score in the separate inattention and hyperactivity-impulsivity domains.

Test statistics are presented in Table 2 for the optimal dichotomous cut-point for each of these six scoring methods, where optimality is defined as minimization of the difference between false positive (FP) and false negative (FN). Method 4 is clearly the best one in this set, yielding the highest total classification accuracy (96·2%), OR (73·4), κ (0·58), and close to the highest AUC (0·77). While compared to Method 4, Method 3 has a marginally higher sensitivity (60·2% *v.* 56·3%) and AUC (78 *v.* 77). This is achieved at the expense of a lower specificity (96·3% *v.* 98·3%), resulting in lower total classification accuracy (94·5% *v.* 96·2%) and a dramatically lower OR (39·5 *v.* 73·4) for Method 3 than Method 4. Based on these considerations and the previously specified optimality criterion, Method 4 was selected as the optimal simple scoring method of the ASRS.

Table 3. *Polychotomous stratification of the optimally scored 18-question ASRS*

	Stratum categorizations					
	0–3		4–8		9–18	
I. Test statistics*						
Sensitivity (S.E.)	13·4%	(6·9)	30·3%	(8·1)	56·3%	(8·8)
1 – Specificity (S.E.)	70·5%	(8·0)	27·8%	(8·0)	1·7%	(0·7)
II. Positive predictive values and standard errors at plausible values of P_P†						
$P_P = 0{\cdot}01$	0·2	(0·1)	1·1	(0·5)	24·8	(8·5)
$P_P = 0{\cdot}03$	0·6	(0·3)	3·3	(1·4)	50·3	(11·4)
$P_P = 0{\cdot}06$	1·2	(6·9)	6·5	(8·1)	67·6	(8·8)
$P_P = 0{\cdot}09$	1·8	(6·9)	9·7	(8·1)	76·4	(8·8)
$P_P = 0{\cdot}12$	2·5	(1·4)	13·0	(4·9)	81·7	(6·8)

* See the text for definitions of the test statistics.
† P_P is the population prevalence of ADHD in a hypothetical population.

In order to explore the usefulness of a polychotomous classification, the optimally-scored ASRS was collapsed into strata so that the probability of being classified as a clinical case did not differ meaningfully across cells within each stratum, but did differ meaningfully across strata. Three strata were found to meet these criteria, corresponding to scale scores in the range 0–3, 4–8, and 9–18. Part I of Table 3 shows the sensitivity and specificity for each stratum. Part II presents estimates of Positive predictive values (PPV) for plausible values of population prevalence. As shown in the first row of Part I, 56·3% of adults in the general population who meet clinical criteria for ADHD are in the highest stratum, with 30·3% in the middle stratum and the remaining 13·4% in the lowest stratum. Nearly three-quarters (70·5%) of clinical non-cases are in the lowest ASRS stratum, 27·8% in the middle stratum, and only 1·7% in the highest stratum. Given that best estimates from available studies put the community prevalence of adult ADHD in the range 1–6% (Wender *et al.* 2001) and prevalence in general medical populations as much as twice as high, the PPV of the three strata in population studies would be in the range 0·2–2·5% in the lowest stratum, 1·1–13·0% in the middle stratum, and 24·8–81·7% in the highest stratum, if sensitivity and specificity were constant.

The ASRS screener

As an 18-question screening scale is too long for many purposes, we investigated whether we could develop a useful short-form screener. Stepwise logistic regression was used to make this evaluation, beginning with the selection of the dichotomously coded screening questions that most accurately predicted the clinical syndrome classifications. An inspection of successive changes in AUC as new questions were added to the prediction equation was used to select an optimal number of questions, leading to the conclusion that six questions were the optimal number. The equation with this number of predictors had AUC = 0·95 and $\chi^2_6 = 13{\cdot}9$ ($p = 0{\cdot}031$). We then attempted to select additional predictors that distinguished between responses of very often and other positive responses. No significant predictors of this sort could be found ($\chi^2_3 = 1{\cdot}4$, $p = 0{\cdot}702$ for the three best predictors in this set). We then attempted to select additional predictors that added information on the full 0–4 range of responses, but again found no significant predictors of this sort ($\chi^2_3 = 0{\cdot}2$, $p = 0{\cdot}978$ for the three best predictors in this set).

Based on these results, further analysis focused on six-question screeners that used dichotomous symptom scoring. All-possible-subsets logistic regression analysis showed that several six-question subsets had very similar values of AUC and χ^2 in predicting clinical syndrome classifications. The concordance of these scales with clinical syndrome classifications was assessed for all logically possible dichotomizations. Test statistics showed one of these to be marginally superior to the rest in that it had close to the highest sensitivity (68·7%), the highest specificity (99·5%), and the highest

Table 4. *Polychotomous stratification of the optimally scored six-question ASRS screener*

	Stratum categorizations					
	0–1		2–3		4–6	
I. Test statistics*						
Sensitivity (S.E.)	4·3%	(2·9)	27·0%	(8·0)	68·7%	(8·2)
1 – Specificity (S.E.)	74·8%	(6·6)	24·7%	(6·5)	0·5%	(0·3)
II. Positive predictive values and standard errors at plausible values of P_P†						
$P_p = 0{\cdot}01$	0·1	(0·0)	1·1	(0·4)	56·8	(14·1)
$P_p = 0{\cdot}03$	0·2	(0·1)	3·3	(1·3)	80·1	(9·2)
$P_p = 0{\cdot}06$	0·4	(0·2)	6·5	(2·5)	89·3	(5·5)
$P_p = 0{\cdot}09$	0·6	(0·4)	9·8	(3·7)	92·8	(3·8)
$P_p = 0{\cdot}12$	0·8	(0·5)	13·0	(4·7)	94·7	(2·9)

* See the text for definitions of the test statistics.
† P_P is the population prevalence of ADHD in a hypothetical population.

overall concordance as indicated by total classification accuracy (97·9%), OR (414·1), κ (0·76), and AUC (0·84). Detailed results of these comparisons are included in the appendix materials posted at the website mentioned earlier in the paper.

As with the optimal 18-question version of the ASRS, the optimal six-question ASRS screener was collapsed into three strata (0–1, 2–3, and 4–6). Sensitivities and specificities for the strata are presented in Part I of Table 4. Approximately two-thirds (68·7%) of clinical cases scored in the highest stratum, with only 4·3% in the lowest stratum. About three-quarters of clinical non-cases (74·8%), in comparison, scored in the lowest stratum and only 0·5% in the highest stratum. Part II of the table shows that PPVs of the three strata are in the range 0·1–0·8 for the lowest stratum, 1·1–3·0 for the middle stratum, and 56·8–94·7 for the highest stratum for plausible values of population prevalence. Comparing these PPVs with the distribution of sensitivities in Part I of the table shows that about one-third of clinical cases (i.e. those in the lowest and middle strata) would be missed with this screen, while two-thirds in the highest stratum would be classified as having a very high probability of being cases.

ASRS discrimination among screener positives

Comparison of the results in Tables 3 and 4 shows that the six-question screener outperforms the full 18-question ASRS. However, two further results suggest that the full ASRS might nonetheless be useful among people who are positive on the screener. First, the percent of clinical cases screening positive on a dichotomous version of the full ASRS in which positives are defined as those scoring 11–18 (49·1%, with an S.E. of 11·0%) is significantly higher than the percent of clinical non-cases screening positive on the same dichotomy (15·2%, with a S.E. of 12·4%, $z = 2{\cdot}0$, $p = 0{\cdot}043$), documenting that administration of the full ASRS can significantly improve classification of true cases among people who are positive on the six-question screener. Second, a substantial Pearson correlation ($r = 0{\cdot}43$, $p < 0{\cdot}001$) exists between a version of the ASRS that sums responses to the 18 ASRS questions using the full 0–4 response scale (generating a scale with a 0–72 range) and the scale of current clinical symptom severity, documenting that repeat administration of the full ASRS might be useful in charting clinical improvement among cases in treatment.

DISCUSSION

It is important to note that no data have ever been published on the validity of the clinical interview used as the gold standard in this study even though it has become the standard in clinical studies of adult ADHD (Spencer *et al.* 1995, 1998, 2001; Michelson *et al.* 2003). To the extent that it imperfectly operationalizes the DSM-IV criteria, the validity of the ASRS might be underestimated. Another important boundary condition is that we developed the scale in an explicit attempt to equalize the

number of false positives and false negatives. Optimization rules that put different weights on false positives and false negatives might have led to different scoring rules or different questions being selected (Kramer, 1992).

A potential limitation of the study design is that the self-report questions were administered after the completion of the clinical assessment, possibly resulting in some respondents becoming more sensitized to their symptoms and responding to the ASRS differently than they would have otherwise. This problem could be resolved in future face-to-face clinical reappraisal studies by having respondents self-administer a paper and pencil version of the ASRS before blind administration of the clinical interview. Another limitation regarding validity is that all data are obtained from respondents rather than also from informants.

Methodological studies comparing adult self-reports *versus* informant reports of ADHD symptoms generally show the same pattern of disagreement as in studies of child self-reports *versus* informant reports (Jensen *et al.* 1999); namely, that informants report higher symptoms than respondents (Gittelman & Mannuzza, 1985; Zucker *et al.* 2002). This suggests that both the clinical interviews and the ASRS results might be conservative. It is important to note, however, that the one adult self-*versus*-informant ADHD symptom comparison study that was carried out in a non-clinical sample found fairly strong associations between the two reports and no self-informant difference in reported symptom severity (Murphy & Schachar, 2000).

An additional limitation is that a single set of scoring rules was presented even though the optimal thresholds and appropriate values of PPV might differ as a function of gender, educational status, marital status, or other known correlates of adult ADHD. No attempt was made to generate subsample scoring rules, however, based on the clinical reappraisal sample being too small for powerful subsample analysis. This small sample size also raises concerns about the generalizability of the results. This is especially true for the ASRS screener, which was developed based on the use of stepwise regression analysis and might have capitalized on chance in selecting items. It is worth noting, in this regard, that the variation in the symptom-level concordance between the ASRS and the clinical ratings could also have been taken into consideration in scoring by including question-level weights, such as those generated in a logistic regression analysis that regressed the dichotomous clinical syndrome classifications on the 18 separate ASRS questions. We decided against this, however, based on concern about over-fitting the data.

Within the context of these limitations, the results suggest that the six-question ASRS screener is a very good tool for reproducing the overall clinical evaluations made by carefully trained and closely monitored clinical interviewers. The three-stratum version of the scale, in particular, has excellent concordance with blind clinical diagnoses. This means that the transformation of the scale's stratum classifications into individual-level predicted probabilities of clinical diagnoses can be used in general population epidemiological surveys to generate an outcome variable that will be a good surrogate for clinical syndrome classifications.

In light of the evidence that the six-question ASRS screener out-performs the full ASRS, a question can be asked whether the latter has any value. As noted in the discussion of limitations, our scoring of the full ASRS was based on an unweighted summation of responses. A weighted version of the full scale might prove to have much greater concordance with clinical syndrome classifications than the screener, although it would be necessary to have a cross-validation sample to investigate this possibility rigorously. Furthermore, we showed that the full ASRS both refines prediction of the clinical classification among respondents who are positive on the six-question screener and correlates significantly with clinician-rated overall symptom severity in this same subsample.

The six-question ASRS screener, in comparison, seems to hold more promise than the full ASRS for clinical screening purposes. As shown in Table 4, over two-thirds of clinical cases screen positive on the six-question screener compared to an extremely low proportion of non-cases (0·5%), resulting in a high proportion of screened positives being true cases under all plausible assumptions about the population prevalence of the disorder. The situation with the roughly one-third of clinical cases who are negative on the six-question ASRS screener is

also important to consider. None of the scoring rules we considered was able to generate a stratum that could reliably distinguish these cases from non-cases. Not surprisingly, these screened negative clinical cases had an average clinical symptom severity score lower than clinical cases that screened positive (1·5 *v.* 1·8, $z=1{\cdot}7$, $p=0{\cdot}091$). More detailed analyses explored whether we could capture these false negatives by using information in the remaining 12 ASRS questions or in the additional 11 questions that were thought by the clinical experts to be other common expressions of adult ADHD. No evidence was found that the screening scale could be improved by using this additional information.

The probability of a person with a given score on a screening scale meeting criteria for a disorder (i.e. the PPV at that point on the scale) will have the same expected value in a given population as in a calibration sample only if the calibration sample is representative of that population. That is why estimates of PPV were reported for a range of plausible prevalence values. Importantly, we found that the PPV of the highest stratum in the ASRS screener is quite high even under the assumption of an implausibly low prevalence. We also found that the PPV of the middle stratum in the ASRS screener is quite low even under the assumption of a high prevalence. These results tell us that patients who screen into the highest stratum of the ASRS screener in primary care samples should routinely be considered likely cases who warrant further evaluation, while primary care patients who screen into the middle stratum should only be evaluated further when there is other evidence to suggest that they might be cases.

Uncertainty about PPV is of considerably more importance, in comparison, in epidemiological surveys that focus on segments of the population that cannot be considered representative of the total USA population. Although it is conventional to use standardized cut-points in such surveys, these can lead to substantial error in estimating prevalence and correlates. A less biased approach is to generate expected stratum distributions for all logically possible values of population prevalence (noting that the expected sample proportion in a given stratum is the sum of the products of prevalence times sensitivity and the additive inverse of prevalence times the additive inverse of specificity) based on the sensitivities and specificities reported in Table 4 and to use maximum-likelihood comparisons of these theoretical distributions with observed stratum distributions in the sample to select the most likely prevalence in the population from which the sample was selected. Once this maximum-likelihood prevalence estimate is obtained, individual-level predicted probabilities can be calculated easily (Guyatt & Rennie, 2001).

Besides using the six-question ASRS screener in epidemiological surveys and in primary care screening, the good results about the precision of the screener and the fact that it can be self-administered easily and quickly (less than 2 minutes) might make it a useful secondary measure to include in clinical studies. This could be a useful complement to the dimensional clinical assessments of ADHD symptom severity typically used in such studies to define a lower-bound severity threshold that distinguishes community cases from non-cases (i.e. the highest *versus* middle strata in the ASRS screener). The use of the ASRS screener in clinical studies would also provide a useful crosswalk between clinical research and community epidemiological research by allowing a comparison of the severity distribution between community and clinical cases. The absence of such comparative data has restricted our ability to interpret the clinical significance of categorical prevalence estimates of most mental disorders in community epidemiological studies up to now (Kessler *et al.* in press). The inclusion of identical short dimensional assessments of adult ADHD in both clinical and community studies would be a useful step in the direction of addressing this important problem for this heretofore understudied disorder.

ACKNOWLEDGEMENTS

The National Comorbidity Survey Replication (NCS-R) is supported by the US National Institute of Mental Health (U01-MH60220) with supplemental support from the US National Institute of Drug Abuse, the Substance Abuse and Mental Health Services Administration, and the Robert Wood Johnson Foundation (Grant no. 044780). Collaborating investigators include Ronald C. Kessler (Principal

Investigator, Harvard Medical School), Kathleen Merikangas (Co-Principal Investigator, NIMH), Doreen Koretz (Co-Principal Investigator, Harvard University), William Eaton (The Johns Hopkins University), Jane McLeod (Indiana University), Mark Olfson (Columbia University College of Physicians and Surgeons), Harold Pincus (University of Pittsburgh), Phillip Wang (Harvard Medical School), Kenneth Wells (UCLA), and Elaine Wethington (Cornell University).

Additional support for the ADHD screening scale validation re-interviews was provided by an unrestricted educational grant from the Eli Lilly Company. The WMH-CIDI Advisory Group for adult ADHD includes Lenard Adler (New York University Medical School), Russell Barkley (Medical College of South Carolina), Joseph Biederman (Massachusetts General Hospital and Harvard Medical School), Keith Conners (Duke University Medical School), Stephen Faraone (Massachusetts General Hospital and Harvard Medical School), Laurence Greenhill (New York State Psychiatric Institute), Molly Howes (Harvard Medical School), Ronald Kessler (Harvard Medical School), Thomas Spencer (Massachusetts General Hospital) and T. Bedirhan Ustun (World Health Organization).

The authors thank the other members of the advisory group for helpful comments on this paper. All NCS-R instruments are posted at http://www.hcp.med.harvard.edu/ncs.

DECLARATION OF INTEREST

L. Adler, S. Faraone, R. C. Kessler and T. Spencer have all served as paid consultants of Eli Lilly and Company. K. Secnik is an employee of Eli Lilly and Company.

REFERENCES

Barkley, R. A. (1995). ADHD behavior checklist for adults. *ADHD Report* **3**, 16.

Barkley, R. A., Fischer, M., Smallish, L. & Fletcher, K. (2002). The persistence of attention-deficit/hyperactivity disorder into young adulthood as a function of reporting source and definition disorder. *Journal of Abnormal Psychology* **111**, 279–289.

Biederman, J., Mick, E. & Faraone, S. V. (2000). Age-dependent decline of symptoms of attention deficit hyperactivity disorder: impact of remission definition and symptom type. *American Journal of Psychiatry* **157**, 816–818.

Bird, H. R., Canino, G. & Rubio-Stipec, M. (1988). Estimates of the prevalence of childhood maladjustments in a community survey in Puerto Rico. *Archives of General Psychiatry* **45**, 1120–1126.

Brown, T. E. (1996). *Brown Attention Deficit Disorder Scales*. Psychological Corporation: San Antonio, CA.

Conners, C. K., Erhardt, D. & Sparrow, E. P. (1998). *Conners CAARS-Self-report: Long Version* (*CAARS-S:L*). Multi-Health Systems: North Totawanda, NY.

DuPaul, G. J., Power, T. J., Anastopoulos, A. D. & Reid, R. (1998). *ADHD Rating Scale-IV: Checklists, Norms, and Clinical Interpretation*. Guilford: New York.

Gittelman, R. & Mannuzza, S. (1985). Diagnosing ADD-H in adolescents. *Psychopharmacology Bulletin* **21**, 237–242.

Guyatt, G. & Rennie, D. (2001). *User's Guide to the Medical Literature: a Manual for Evidence-based Clinical Practice*. AMA Press: Chicago, IL.

Heiligenstein, E., Conyers, L. M., Berns, A. R., Miller, M. A. & Smith, M. A. (1998). Preliminary normative data on DSM-IV attention deficit hyperactivity disorder in college students. *Journal of the American College Health Association* **46**, 185–188.

Jensen, P. S., Rubio-Stipec, M., Canino, G., Bird, H. R., Dulcan, M. K., Schwab-Stone, M. E. & Lahey, B. B. (1999). Parent and child contributions to diagnosis of mental disorder: are both informants always necessary? *Journal of the American Academy of Child and Adolescent Psychiatry* **38**, 1569–1579.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., Walters, E. E. & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine* **32**, 959–976.

Kessler, R. C., Berglund, P., Chiu, W. T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B. E., Walters, E. E., Zaslavsky, A. & Zheng, H. (in press). The US National Comorbidity Survey Replication (NCS-R): an overview of design and field procedures. *International Journal of Methods in Psychiatric Research*.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., Rush, A. J., Walters, E. E. & Wang, P. S. (2003). The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association* **289**, 3095–3105.

Kessler, R. C. & Ustun, T. B. (2000). The World Health Organization World Mental Health 2000 Initiative. *Hospital Management International*, 195–196.

Kramer, H. C. (1992). *Evaluating Medical Tests: Objective and Quantitative Guidelines*. Sage Publications: Newbury Park, CA.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Mannuzza, S., Klein, R. G., Bessler, A., Malloy, P. & LaPadula, M. (1993). Adult outcome of hyperactive boys. Educational achievement, occupational rank, and psychiatric status. *Archives of General Psychiatry* **50**, 565–576.

Mannuzza, S., Klein, R. G., Bessler, A., Malloy, P. & LaPadula, M. (1998). Adult psychiatric status of hyperactive boys grown up. *American Journal of Psychiatry* **155**, 493–498.

Mehringer, A. M., Downey, K. K., Schuh, L. M., Pomerleau, C. S., Snedecor, S. M. & Schbiner, H. (2002). The assessment of hyperactivity and attention (AHA): development and preliminary validation of a brief self-assessment of adult ADHD. *Journal of Attention Disorders* **5**, 223–231.

Menkes, M. M., Rowe, J. S. & Menkes, J. H. (1967). A twenty-five-year follow-up study on the hyperkinetic child with minimal brain dysfunction. *Pediatrics* **39**, 393–399.

Michelson, D., Adler, L., Spencer, T., Reimherr, F. W., West, S. A., Allen, A. J., Kelsey, D., Wernicke, J., Dietrich, A. & Milton, D. (2003). Atomoxetine in adults with ADHD: two randomized, placebo-controlled studies. *Biological Psychiatry* **53**, 112–120.

Murphy, K. & Barkley, R. A. (1996). Attention deficit hyperactivity disorder in adults: comorbidities and adaptive impairments. *Comprehensive Psychiatry* **37**, 393–401.

Murphy, P. & Schachar, R. (2000). Use of self-ratings in the assessment of symptoms of attention deficit hyperactivity disorder in adults. *American Journal of Psychiatry* **157**, 1156–1159.

O'Donnell, J. P., McCann, K. K. & Pluth, S. (2001). Assessing adult ADHD using a self-report symptom checklist. *Psychological Reports* **88**, 871–881.

Pary, R., Lewis, S., Matuschka, P. R., Rudzinskiy, P., Safi, M. & Lippmann, S. (2002). Attention deficit disorder in adults. *Annals of Clinical Psychiatry* **14**, 105–111.

Peirce, J. C. & Cornell, R. G. (1993). Integrating stratum-specific likelihood ratios with the analysis of ROC curves. *Medical Decision Making* **13**, 141–151.

Robins, L. N., Cottler, L., Bucholz, K. & Compton, W. (1995). *Diagnostic Interview Schedule for DSM-IV*. Washington University: St. Louis, MO.

Robins, L. N., Helzer, J. E., Croughan, J. L. & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristics and validity. *Archives of General Psychiatry* **38**, 381–389.

Robins, L. N., Wing, J., Wittchen, H.-U., Helzer, J. E., Babor, T. F., Burke, J. D., Farmer, A., Jablenski, A., Pickens, R., Regier, D. A., Sartorius, N. & Towle, L. H. (1988). The Composite International Diagnostic Interview: an epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry* **45**, 1069–1077.

SAS Institute (1999). *SAS User's Guide, release* 8. SAS Institute Inc.: Cary, NC.

Shekim, W. O., Kashani, J. & Beck, N. (1985). The prevalence of attention deficit disorders in a rural Midwestern community sample of nine-year-old children. *Journal of the American Academy of Child Psychiatry* **24**, 765–770.

Spencer, T., Biederman, J., Wilens, T., Faraone, S., Prince, J., Gerard, K., Doyle, R., Parekh, A., Kagan, J. & Bearman, S. K. (2001). Efficacy of a mixed amphetamine salts compound in adults with attention-deficit/hyperactivity disorder. *Archives of General Psychiatry* **58**, 775–782.

Spencer, T., Biederman, J., Wilens, T., Prince, J., Hatch, M., Jones, J., Harding, M., Faraone, S. V. & Seidman, L. (1998). Effectiveness and tolerability of tomoxetine in adults with attention deficit hyperactivity disorder. *American Journal of Psychiatry* **155**, 693–695.

Spencer, T., Wilens, T., Biederman, J., Faraone, S. V., Ablon, J. S. & Lapey, K. (1995). A double-blind, crossover comparison of methylphenidate and placebo in adults with childhood-onset attention-deficit hyperactivity disorder. *Archives of General Psychiatry* **52**, 434–443.

Weiss, G., Hechtman, L., Milroy, T. & Perlman, T. (1985). Psychiatric status of hyperactives as adults: a controlled prospective 15-year follow-up of 63 hyperactive children. *Journal of the American Academy of Child Psychiatry* **24**, 211–220.

Wender, P. H. (1998). Attention-deficit hyperactivity disorder in adults. *Psychiatric Clinics of North America* **21**, 761–774.

Wender, P. H., Wolf, L. E. & Wasserstein, J. (2001). Adults with ADHD. An overview. *Annals of the New York Academy of Sciences* **931**, 1–16.

West, S. L., Muslow, M. & Arredondo, R. (2003). Factor analysis of the Attention Deficit Scales for Adults (ADSA) with a clinical sample of outpatient substance abusers. *American Journal of Addiction* **12**, 159–165.

Wilens, T. E., Biederman, J. & Spencer, T. J. (2002). Attention deficit/hyperactivity disorder across the lifespan. *Annual Review of Medicine* **52**, 113–131.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag: New York.

Zucker, M., Morris, M. K., Ingram, S. M., Morris, R. D. & Bakeman, R. (2002). Concordance of self- and informant ratings of adults' current and childhood attention-deficit/hyperactivity disorder symptoms. *Psychological Assessment* **14**, 379–389.