The Neuropsychological Effects of Obstructive Sleep Apnea: A Meta-Analysis of Norm-Referenced and Case-Controlled Data

Dean W. Beebe, PhD; Lisa Groesz, BA; Carolyn Wells, MA; Alisha Nichols, BA; Kevin McGee, BA

1Cincinnati Children’s Hospital Medical Center and the University of Cincinnati, Cincinnati, Ohio; 2Xavier University, Cincinnati, Ohio

Study Objectives: The research literature on the neuropsychological effects of obstructive sleep apnea (OSA) has yielded seemingly contradictory findings, and narrative reviews of this literature are prone to interpretive errors. We used sophisticated meta-analytic models to minimize such errors, with the goal of clarifying the effect of OSA on neuropsychological functioning.


Participants: We reviewed studies of neuropsychological functioning among adults with untreated OSA. Twenty-five studies met review criteria, representing 1092 patients with OSA and 899 healthy controls.

Measurements and Results: Two sets of effect sizes were generated. One compared OSA group means against those of healthy controls in case-controlled studies. The other compared all OSA group means against published normative data. Within each data set, 10 neuropsychological outcome domains were coded. In both data sets, untreated OSA was found to have a negligible impact on intellectual and verbal functioning but a substantial impact upon vigilance and executive functioning. Data were mixed with regard to visual and motor functioning; post hoc inspection of the data suggested that tests of fine-motor coordination or drawing were more sensitive to OSA than were tests of fine-motor speed or visual perception. Data were also mixed with regard to memory functioning, probably related to methodologic differences across studies.

Conclusions: Etiologic models should emphasize mechanisms known to affect vigilance, executive functioning, and motor coordination but not intelligence, verbal functioning, or visual perception. Clinicians should be alert to OSA symptoms in patients with declines in vigilance, executive functioning, or coordination.


INTRODUCTION

IT IS WELL RECOGNIZED THAT OBSTRUCTIVE SLEEP APNEA (OSA) IS ASSOCIATED WITH DAYTIME SLEEPINESS THAT RESULTS IN OCCUPATIONAL DEFICITS AND AN INCREASED RISK OF AUTOMOBILE ACCIDENTS.1-3 Though less publicized, OSA’s effects on cognition pose additional challenges to the adaptive functioning of patients.4 These neuropsychological effects can not be subsumed under the term sleepiness because doing so would require expansion of the already multidimensional construct of sleepiness beyond reasonable limits of conceptual specificity.4 Moreover, some researchers have found that neuropsychological deficits correlate better with polysomnographic sleep data than with self-reported or objectively measured sleepiness.5-6 Indeed, such deficits may persist despite treatment-related resolution of daytime sleepiness.7-9 As such, the neuropsychological effects of OSA are important in their own right.

Unfortunately, the current literature can be confusing. Studies have varied widely in methods. The use of numerous outcome measures without attention to the accumulation of alpha across analyses has resulted in Type I “false positive” errors. Small sample sizes in clinical studies have further promoted Type II “false negative” errors due to inadequate statistical power. Analogous errors can occur when reviewing the literature. Reviewers can make a Type I error when they accept any single study’s statistically significant finding without consideration of other studies that present contrary data. A related error occurs when all statistically significant effects are treated equally, resulting in the conclusion that many functions are equivalently impaired when, in fact, a comparison of effect sizes is required to test such a conclusion. Type II error can occur when a reviewer uses a simple “vote-count” tally of studies with significant and nonsignificant findings on a given variable. Counterintuitively, this method may increase Type II error as the number of reviewed articles increases.10

The empirical review strategy of meta-analysis can often minimize these errors by quantitatively pooling findings across multiple studies. Type I error is minimized when this pooling “washes out” sample idiosyncrasies that can result in spurious findings in small-sampled individual studies. Analyses are based upon continuous effect sizes, not binary statistical significance decisions. Type II error is minimized as small samples are pooled into a much larger combined sample. Engelman and colleagues11 recently applied meta-analytic techniques to published case-controlled studies on the cognitive effects of OSA. They reported deficits in attention, psychomotor functioning, and executive functioning (eg, mental flexibility, planning, problem-solving, and “online” mental manipulation of information). Somewhat smaller effects were reported in learning and memory. Their publication provided an important summary of the literature, and its focus on published case-controlled research may have maximized the quality of the reviewed research. However, its narrow focus reduced the pool of reviewed studies to 5, restricting the domains of functioning that could be analyzed and preventing statistical tests of potential effect moderators, such as disease severity. Moreover, the reliance on published literature prevented examination of possible publication bias that can result in the disproportionate representation of null findings in unreported research.

The present review was intended to build upon that presented by Engelman and colleagues.11 Like that review, we used quantitative meta-analytic techniques to minimize Type I and Type II errors. However, we evaluated all available quantitative data on the neuropsychological effects of OSA, including both case-controlled and uncontrolled studies (eg, baseline data from treatment studies). This resulted in two complementary sets of effect sizes: (a) a comparison of OSA patients to within-study healthy controls (hereafter referred to as the control-referenced data set) and (b) a comparison of OSA patients to published normative...
data (non-referenced data set). Including data from uncontrolled studies dramatically increased pooled sample sizes, adding statistical power and allowing for examination of the moderating effects of research design (case-controlled vs. uncontrolled methods) and disease severity. We also included unpublished dissertations in the analyses to assess possible medication effects of publication status or bias. Because a large number of studies were eligible for review, we were able to examine an unprecedented breadth of neuropsychological outcome domains. To provide the most technically accurate representation of findings, we used advanced random and mixed-effects meta-analytic models (described later), which, to our knowledge, have not been applied in previous OSA reviews. The technical challenges associated with this work required us to limit our scope to the neuropsychological effects of untreated OSA; treatment effects were beyond the scope of this review.

METHOD

Sample Selection

Sample selection proceeded through several phases. First, a preliminary list of published references was generated via PubMed and PsychINFO databases, cross-referencing “apnea OR snoring” with “cognitive

Table 1—Meta-analyzed studies and descriptive characteristics

<table>
<thead>
<tr>
<th>Citation</th>
<th>Year</th>
<th>Sample Source</th>
<th>Control Demographics</th>
<th>OSA Demographics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td>Age</td>
</tr>
<tr>
<td>A Bailey (Sample 1)</td>
<td>1993</td>
<td>R</td>
<td>D</td>
<td>10</td>
</tr>
<tr>
<td>B Bailey (Sample 2)</td>
<td>1993</td>
<td>R</td>
<td>D</td>
<td>7</td>
</tr>
<tr>
<td>C Barbé et al. 1998</td>
<td>R</td>
<td>A</td>
<td>60</td>
<td>47</td>
</tr>
<tr>
<td>D Bardwell et al (pre-placebo)</td>
<td>2001</td>
<td>R</td>
<td>A</td>
<td>16</td>
</tr>
<tr>
<td>E Bardwell et al (pre-CPAP)</td>
<td>2001</td>
<td>R</td>
<td>A</td>
<td>20</td>
</tr>
<tr>
<td>F Bedard et al (moderate OSA)</td>
<td>1991</td>
<td>R</td>
<td>A</td>
<td>50</td>
</tr>
<tr>
<td>G Bedard et al (severe OSA)</td>
<td>1991</td>
<td>R</td>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td>H Borak et al. 1996</td>
<td>R</td>
<td>A</td>
<td>20</td>
<td>46</td>
</tr>
<tr>
<td>I Cammermeyer 1991</td>
<td>R</td>
<td>D</td>
<td>11</td>
<td>51</td>
</tr>
<tr>
<td>J Engleman et al. 2000</td>
<td>R</td>
<td>A</td>
<td>98</td>
<td>47</td>
</tr>
<tr>
<td>K Findley et al. 1991</td>
<td>R</td>
<td>A</td>
<td>21</td>
<td>59</td>
</tr>
<tr>
<td>L Findley et al. 1995</td>
<td>R</td>
<td>A</td>
<td>10</td>
<td>48</td>
</tr>
<tr>
<td>M Findley et al. 1999</td>
<td>R</td>
<td>A</td>
<td>14</td>
<td>43</td>
</tr>
<tr>
<td>N Froehling 1991</td>
<td>R</td>
<td>D</td>
<td>41</td>
<td>47</td>
</tr>
<tr>
<td>O Greenberg et al. 1987</td>
<td>R</td>
<td>A</td>
<td>14</td>
<td>44</td>
</tr>
<tr>
<td>P Ingram et al. 1994</td>
<td>N</td>
<td>A</td>
<td>43</td>
<td>62</td>
</tr>
<tr>
<td>Q Kim et al. 1997</td>
<td>N</td>
<td>A</td>
<td>642</td>
<td>199</td>
</tr>
<tr>
<td>R Kloufet al. 1987</td>
<td>R</td>
<td>A</td>
<td>11</td>
<td>49</td>
</tr>
<tr>
<td>S Kribbs et al. 1993</td>
<td>R</td>
<td>A</td>
<td>45</td>
<td>46</td>
</tr>
<tr>
<td>T Lojander et al. 1999</td>
<td>R</td>
<td>A</td>
<td>49</td>
<td>48</td>
</tr>
<tr>
<td>U Merton 1991</td>
<td>R</td>
<td>D</td>
<td>25</td>
<td>49</td>
</tr>
<tr>
<td>V Monasterio 2001</td>
<td>R</td>
<td>A</td>
<td>59</td>
<td>54</td>
</tr>
<tr>
<td>W Monasterio (before conservative tx)</td>
<td>2001</td>
<td>R</td>
<td>A</td>
<td>66</td>
</tr>
<tr>
<td>X Naegele 1995</td>
<td>R</td>
<td>A</td>
<td>17</td>
<td>49</td>
</tr>
<tr>
<td>Y Redline 1997</td>
<td>R</td>
<td>A</td>
<td>20</td>
<td>49</td>
</tr>
<tr>
<td>Z Sloan 1990</td>
<td>R</td>
<td>D</td>
<td>19</td>
<td>45</td>
</tr>
<tr>
<td>AAA Sloan 1990</td>
<td>R</td>
<td>D</td>
<td>19</td>
<td>45</td>
</tr>
<tr>
<td>BB Valencia et al. Flores et al. 1996</td>
<td>R</td>
<td>A</td>
<td>37</td>
<td>49</td>
</tr>
<tr>
<td>CC Verslaet et al. 1996</td>
<td>R</td>
<td>A</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>DD Walker 1990</td>
<td>R</td>
<td>D</td>
<td>30</td>
<td>47</td>
</tr>
<tr>
<td>EE Walker (“tx rejecters”)</td>
<td>1990</td>
<td>R</td>
<td>D</td>
<td>9</td>
</tr>
</tbody>
</table>

Notes: Citations are listed alphabetically by first author and assigned letters for ease of presentation in Table 2. The sample column indicates whether the obstructive sleep apnea (OSA) subjects were clinic referred (R) or derived from a screened community sample (N). Source refers to the publication status of the citation—non-peer-reviewed (D, dissertation) versus peer-reviewed (A, published article). Demographic data refer to each sample as a whole, including sample size N, mean age, percent male, mean years of formal education (Ed), and mean body mass index (BMI, kg/m²). For the OSA groups, mean apnea index (AI) and apnea-hypopnea index (AHI) refer to the average number of obstructive apneas or apneas+hypopneas per hour of sleep, respectively, for each sample. CPAP = continuous positive airway pressure.

(1) Childhood age. Because of a paucity of data from children and differences in disease characteristics from childhood to adulthood, we excluded 4 studies with subjects under age 18.

(2) Biased sample. Four studies were excluded because they reported on patients who were preselected to have extreme scores on a neuropsychological test (eg, patients with dementia).

(3) Subthreshold severity. We excluded 11 studies that reported their clinical (OSA) group only to snore, without polysomnographic evidence of an obstructive apnea-hypopnea index (AHI) greater than 5, or with evidence that some subjects had an AHI less than 5.

(4) Lack of baseline data. Two studies were excluded that reported only posttreatment outcome.

(5) Central apnea. We excluded 2 samples that included patients with primarily central apneas.

(6) Lack of neuropsychological tests. Fourteen studies were excluded that did not report central tendency statistics for any neuropsychological test or reliance was made on an overly broad neuropsychological measure such as a dementia screening.

(7) Insufficient information to locate norms. Two studies were excluded entirely because of insufficient test documentation (eg, nonstandard measure or administration procedure, insufficient measure citation), but we included numerous studies that had incomplete documentation for some tests.

Medically compromised comparison groups (eg, patients with chronic obstructive pulmonary disease) were disregarded; such studies were treated as uncontrolled rather than case controlled. When multiple clinical groups were reported within the same study (eg, 2 groups assigned to different treatment conditions),
these were treated as separate samples. Application of these review criteria resulted in 25 eligible research studies, representing 32 samples, 1092 total subjects with OSA, and 899 healthy controls. Table 1 lists these studies and related descriptive statistics.

**Sample Coding**

Each study’s methods, sample characteristics, and neuropsychological outcomes were double-coded by the first and second authors to minimize transcription errors and to avoid overlooking relevant data. We intentionally avoided subjective ratings, and raters were not blinded to each other’s coding. Any coding disagreements were resolved by consensus. The following methodologic variables were coded: presence or absence of a healthy control group, source of the OSA sample (community vs. clinical), and publication status (unpublished dissertation vs. peer-reviewed published article). Several descriptive characteristics were reported often enough to code across studies: sample size, sample age, percent male, years of formal education, and body mass index (BMI). We attempted to code for ethnicity and sleepiness, but no more than 8 articles provided data on these variables, so they are not presented here.

Two disease-severity indexes were coded: apnea index (AI; obstructive apneas per hour during polysomnogram) and AHI (obstructive apneas plus hypopneas per hour during sleep). Although it would have been optimal to code other sleep and respiratory data (eg, minimum blood oxygen saturation, number of arousals per hour), less than half of the studies reported such data. A rough severity index was derived based upon recommendations set by the American Academy of Sleep Medicine.

Table 2—Coded domains, measures, and samples

<table>
<thead>
<tr>
<th>Domain</th>
<th>Definition</th>
<th>Measures</th>
<th>Norms used to compute effect sizes</th>
<th>Source References from Table 1 (Case-controlled studies are italicized)</th>
<th># of OSA patients in analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal Ability</td>
<td>Basic verbal lexicon, receptive and expressive language fundamentals</td>
<td>WAIS or WAIS-R Verbal IQ or its subtests</td>
<td>(45, 46)</td>
<td>A, B, F, G, H, N, O, P, T</td>
<td>50</td>
</tr>
<tr>
<td>Visual Ability</td>
<td>Visual perceptual accuracy, visual processing, reproducing visual designs</td>
<td>WAIS or WAIS-R Performance IQ or its subtests, Hooper VOT, ROCF Copy, Raven Matrices</td>
<td>(45-49)</td>
<td>A, B, F, G, H, J, O, T, U, V, W, Z, AA, CC</td>
<td>76</td>
</tr>
<tr>
<td>Short-term verbal memory</td>
<td>Ability to repeat back oral information immediately after it was presented</td>
<td>Buschke SRT, WMS Logical Memory or Word Pairs subtests, RAVLT, CVLT Free</td>
<td>(50-55)</td>
<td>A, B, F, G, K, N, O, Q, U, V, W, X, Y, Z, AA, BB, DD, EE</td>
<td>374</td>
</tr>
<tr>
<td>Long-term verbal memory</td>
<td>Ability to repeat back oral information 20-30 minutes after it was presented</td>
<td>Delayed recall trials of the short-term verbal memory measures</td>
<td>(50-54, 58)</td>
<td>A, B, F, G, K, N, O, Q, U, X, Y, Z, AA, BB, DD, EE</td>
<td>374</td>
</tr>
<tr>
<td>Long-term visual memory</td>
<td>Ability to recognize or reproduce visual information 20-30 minutes after seeing it</td>
<td>Delayed recall trials of the short-term visual memory measures</td>
<td>(53, 57, 58)</td>
<td>A, B, F, G, K, O, U, X, Z, AA, DD, EE</td>
<td>143</td>
</tr>
<tr>
<td>Vigilance</td>
<td>Ability to maintain attention over an extended period</td>
<td>Steer Clear, PVT, Continuous Performance Tests</td>
<td>N/A (see Methods)</td>
<td>C, F, G, K, L, M, P, Y, Z, AA</td>
<td>292</td>
</tr>
</tbody>
</table>

**Abbreviations:** OSA = obstructive sleep apnea, WAIS or WAIS-R = Wechsler Adult Intelligence Scale (original or revised), VOT = Visual Organization Test, ROCF = Rey-Osterreith Complex Figure, SRT = Selective Reminding Test, WMS or WMS-R = Wechsler Memory Scale (original or revised), RAVLT = Rey Auditory Verbal Learning Test, CVLT = California Verbal Learning Test, CVMT = Continuous Visual Memory Test, VRT = Visual Reproduction Test, FR = Facial Recognition, COWAT = Controlled Oral Word Association Test, PASAT = Paced Auditory Serial Addition Test, WCST = Wisconsin Card Sorting Test, PVT = Psychomotor Vigilance Test.
Samples with mean AHI greater than 30 or mean AI greater than 20 were classified as "severe." Samples in which respiratory data were reported but neither condition was true, as well as community-recruited samples for which no respiratory data were reported, were classified as "mild to moderate."  

Finally, 10 outcome domains, based upon accepted neuropsychological practice, were coded when present. Operational definitions, studies coded, normative data accessed, and the number of subjects analyzed within each domain are listed in Table 2 (sample sizes varied because no study measured all domains). Though vigilance is known to be sensitive to sleep deprivation and disruption, analyzing this domain posed a logistic challenge. This domain is typically measured with a Continuous Performance Test (CPT), but dozens of CPTs exist, very few of which have published norms. Studies of OSA patients often failed to specify which CPT was used, and the most commonly used CPT in the sleep literature, "Stee Clear," does not have published norms that are independent of case-controlled studies of OSA (Larry Findley, personal communication, October 14, 2002). Consequently, only case-controlled studies were coded for the vigilance domain.

Following Engleman and colleagues, control-referenced effect sizes were derived from case-controlled studies and were defined as the difference between the OSA and control group mean scores, divided by the control group standard deviation. The control group standard deviation (rather than a pooled standard deviation across both groups) was used because variations in disease severity within the clinical group had the potential to inflate outcome variability.

We also constructed a norm-referenced data set based upon all 32 OSA samples (case-controlled and uncontrolled). We attempted to obtain the best possible normative data on instruments used in the clinical groups, first by consulting published normative compendia, and then via a search of the neuropsychological literature. The normative studies we used are shown in Table 2. Selection of norms was based primarily upon psychometric strengths of the normative group (as outlined by Mitrushina), the mean age of each OSA sample, the modal gender in that sample (when norms allowed), and the education of the sample (when norms allowed). When norms required information that was not clear from the source reference (eg, education not reported), we used the following surrogates: male gender, age 50, and 12 to 13 years of education. Norm-referenced effect sizes were defined as the difference between the OSA group mean and the mean score of the appropriate normative group, divided by the standard deviation of the normative group.

Table 3—Weighted mean effect sizes (d) for the norm-referenced data set

<table>
<thead>
<tr>
<th>Domain</th>
<th>Is d ≠ 0?</th>
<th>Homogeneity statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligence</td>
<td>d (95% CI)</td>
<td>Z  p</td>
</tr>
<tr>
<td>Uncontrolled</td>
<td>-.67 (-1.22 - -1.13)</td>
<td>-2.4 .016*</td>
</tr>
<tr>
<td>Case-Controlled</td>
<td>-.11 (-.32 - -.31)</td>
<td>-.5 .619</td>
</tr>
<tr>
<td>Verbal</td>
<td>-.59 (-.32 - -.25)</td>
<td>-3.5 .001***</td>
</tr>
</tbody>
</table>

Table 4—Weighted mean effect sizes (d) for the control-referenced data set

<table>
<thead>
<tr>
<th>Domain</th>
<th>Is d ≠ 0?</th>
<th>Homogeneity statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligence</td>
<td>d (95% CI)</td>
<td>Z  p</td>
</tr>
<tr>
<td>Verbal</td>
<td>.13 (-.40 - -.66)</td>
<td>.5 .627</td>
</tr>
<tr>
<td>Visual</td>
<td>.68 (.04 - 1.3)</td>
<td>2.1 .039*</td>
</tr>
<tr>
<td>ST Verbal Memory</td>
<td>.29 (-.01 - .59)</td>
<td>.9955</td>
</tr>
<tr>
<td>ST Visual Memory</td>
<td>.60 (-.25 - .26)</td>
<td>0.0 .972</td>
</tr>
<tr>
<td>LT Verbal Memory</td>
<td>.52 (-.13 - .91)</td>
<td>.26 .010*</td>
</tr>
<tr>
<td>LT Visual Memory</td>
<td>.14 (-.21 - .50)</td>
<td>.8 .423</td>
</tr>
<tr>
<td>Executive Function</td>
<td>.53 (-.28 - .76)</td>
<td>.4 .003</td>
</tr>
</tbody>
</table>

Statistical Analyses

The most straightforward and common approach to meta-analysis, which appears to have been used by Engleman et al., is the fixed effects model. This model assumes that variation in effect sizes across studies is the result of sampling error or known moderator variables. This assumption may be questioned, however, in light of the varying methods used to recruit, assess, and screen OSA subjects across studies. The alternatives to the fixed effects model include the random and mixed-effects models. The random-effects model considers an additional source of variation: variation in true effect sizes that result from the numerous unknown or unmeasured differences that occur across studies (labeled random effects). The mixed-effects model is a statistical hybrid, in which the effects of specified moderators are considered along with sampling error and random effects. The random and fixed effects models are flexibe; they can accommodate the presence of random effects, yet they yield the same results as a fixed effects model if no random effects are present in a given data set. This conceptual sophistication and flexibility led us to use random and mixed-effects meta-analytic models for this review.

Analyses were conducted using SPSS 10.0 for Windows (SPSS Inc., Chicago, Illinois) and meta-analysis macros for SPSS provided by Lipsey and Wilson. To pool samples of different sizes, initial inverse variance weights within each sample were computed using Lipsey and Wilson’s formula number 3.24 for the control-referenced data set and the square root of each OSA sample size for the norm-referenced data set (David Wilson, personal communication, October 21, 2001). Inverse variance weights place greater emphasis on large sample studies (which are presumed to be less susceptible to sampling error) than on small sample studies.

We first analyzed the norm-referenced data set because it allowed us to test moderator effects. For each of the outcome domains, regression-based weighted mixed-effects models examined 3 potential moderators: study design (case controlled vs. uncontrolled), sample source (non-peer-reviewed vs. peer-reviewed), and disease severity (mild/moderate vs. severe). Because disease severity was not reported for 1 sample, we ran analyses with and without this variable. We had initially intended to consider the source of the OSA sample (clinical vs. community) as a potential moderator, but only 3 community-recruited samples were eligible for review and all 3 fell in the mild-moderate range of...
We anticipated the heterogeneity of effect sizes moderated by control-group presence, sample source, or disease severity. Follow-up weighted random-effects analyses were conducted within each level of the moderator. If the effect size on a given outcome domain was not found to be moderated by any of these 3 variables, the weighted random-effects model was computed based upon the sample as a whole. The weighted random-effects model was then computed on each outcome domain within the control-referenced data set; this data set was too small to conduct moderator analyses.

Finally, within each domain, we assessed the homogeneity of effect sizes for the norm-referenced and control-referenced data sets. This was accomplished with the Q-statistic, which, when significant, indicates greater variability in effect sizes within a given domain than would be expected by sampling error alone. When this occurred, we posthoc explored the data in an effort to identify potential sources of this variability.

RESULTS

Norm-Referenced Data

Within the norm-referenced data set, initial mixed-model regression analyses with study design and publication status as predictors indicated that study design significantly moderated the effect size in 2 domains: intelligence (p = .035) and visual ability (p < .001). Adding disease severity to the predictors reduced the effect of study design on the intelligence domain to nonsignificant levels, but the effect of study design on visual-ability findings remained robust (p = .002). Neither publication status nor disease severity significantly moderated effect size in any analysis (p > .05).

Table 3 provides the norm-referenced, weighted, mean effect sizes across domains, the 95% confidence interval for this mean; a z-score examining the null hypothesis that the mean effect size is 0; and the Q-statistic testing the null hypothesis that the effect sizes across studies were homogeneous. For the domains of intelligence and visual ability, separate analyses were conducted within the case-controlled studies and the uncontrolled studies. As can be seen in Table 3, the case-controlled studies yielded generally higher effect sizes than did uncontrolled studies on these two domains. However, the mean effect sizes in the case-controlled studies failed to reach statistical significance; both p values were greater than .05. In the visual domain, the mean weighted effect size among the case-controlled studies appeared large, but markedly inconsistent findings across studies led to a broad confidence interval and related statistical non-significance. The mean effect sizes in uncontrolled studies were significant in the negative direction for intelligence (d = .67, p = .016), but non-significant for visual functioning.

Within the norm-referenced data set as a whole, the weighted mean effect sizes were small and non-significant in the domains of short-term verbal and visual memory, long-term visual memory, and motor functioning. The verbal domain yielded a significant negative effect (d = -.59, p = .001). In contrast, the long-term verbal memory and executive functioning domains yielded nearly identical mean effect sizes of .52 and .53, respectively, both statistically significant, p values of .01 or less.

Control-Referenced Data

Results for the control-referenced data set are presented in Table 4. To allow direct comparison, weighted mean effect sizes across data sets are plotted in Figure 1. Within the control-referenced data set, the domains of intelligence, verbal ability, and short-term verbal memory yielded small and non-significant mean effect sizes. The executive functioning mean effect size of .73 was significant (p < .001). Unlike the norm-referenced data, the control-referenced data set did not indicate a significant deficit in long-term verbal memory (d = .27, p > .05). Rather, the control-referenced data set reflected prominent vigilance deficits (d = 1.40, p < .001), as well as visual and visual-motor deficits, with significant mean effect sizes noted in visual ability (d = .68, p < .05), motor skills (d = 1.21, p = .001), short-term visual memory (d = .56, p = .001), and long-term visual memory (d = .55, p < .001).

Homogeneity Analyses

A number of Q-statistics were statistically significant in both data sets, even after statistically significant moderators were taken into account. Both the norm-referenced and control-referenced data sets indicated significant heterogeneity of effect sizes within the following domains: verbal and visual short-term and long-term verbal memory, and executive and motor functioning. Significant heterogeneity was also noted in the vigilance domain in the control-referenced data set and in the short-term visual memory domain in the norm-referenced data set. Posthoc visual inspection of scores within the motor and visual domains suggested that some variability was related to the outcome measures used by each study. Within the motor domain, studies that measured fine-motor speed and coordination (pegboard tasks) yielded mean effect sizes of .39 (norm-referenced data set) and 1.94 (control-referenced data set). By contrast, studies that measured only simple finger oscillation speed (finger tapping) yielded lower mean effect sizes of -.83 (norm-referenced data) and .57 (control-referenced data). Within the visual domain, the 3 studies that included drawing tasks, which also rely heavily upon fine-motor coordination, yielded mean effect sizes of 1.46 (norm-referenced data set) and .90 (control-referenced data set). The visual domain effect size in the samples that did not include drawing tasks was much smaller (.23 in the norm-referenced data set, .42 in the control-referenced data set). The pattern of heterogeneity was less clear in the executive, vigilance, short-term visual memory, or short-term or long-term verbal memory domains, though the majority of measures employed likely contributed to the variability across studies.

DISCUSSION

Using statistically appropriate random- and mixed-effects models, this review established the presence and degree of neuropsychological morbidity associated with OSA. Two methods of coding morbidity—morbidity relative to published norms and morbidity relative to healthy controls—yielded results with points of convergence and divergence. Comparison against published norms allowed us to summarize a wide range of neuropsychological effects associated with OSA.
of studies. This increased the pooled sample sizes, provided for a broad scope of outcome domains, and allowed for evaluation of 3 potential moderators: presence or absence of a control group, publication or editorial status, and a rough index of disease severity. The main findings are summarized as separate points below.

General Intelligence and Verbal Ability are Typically Unaffected by OSA

Neither the norm-referenced nor control-referenced data sets suggested significant pathology in overall intelligence or basic language capacities. In fact, mean performance of OSA patients on intellectual and verbal measures was better than published norms. To some degree, this counter-intuitive finding may be related to selection bias; insofar as unusually intelligent or verbal individuals are more likely to volunteer for research, this would inflate scores within these domains. A second potential contributor to this finding is a well-replicated (if not well-understood) psychometric phenomenon in which the norms for intelligence tests “drift” over time. On average, today’s mean score on each of these tests in the general population tends to be about one fifth of a standard deviation unit higher than the mean score on the same test 10 years ago. It is noteworthy that most of the OSA studies used tests of intelligence and verbal ability that had been normed at least a decade earlier, some several decades earlier. This “normative drift” would not impact the control-referenced data set, as both the OSA and healthy subjects received the same instrument at the same time. Given this, it is not surprising that the control-referenced data set resulted in effect sizes near 0.

Vigilance is Markedly Affected by OSA

The domain of vigilance displayed a very large effect size (1.40 in control-referenced analyses), indicating that OSA markedly impairs the ability to sustain attention for extended periods. This finding has obvious implications for the driving and occupational functioning of individuals with untreated OSA and is consistent with Engleman and colleagues’ recent literature review. Impressively, every study that met inclusion criteria and that used a measure of vigilance yielded at least a moderate effect, despite between-study differences in the vigilance test used and in sample acquisition and composition. The magnitude and robustness of this finding supports routine screening of vigilance in patients with suspected OSA, as well as consideration of possible vigilance impairments when making treatment decisions. However, clinicians and researchers alike are cautioned to attend to the psychometric aspects of the vigilance tasks they use, especially to the minimal normative data that are available for most of these tasks.

Executive Functioning is Substantially Affected by OSA

The domain of executive functioning displayed a moderate to large mean effect size (.53 in norm-referenced analyses, .73 in control-referenced analyses). On a theoretical level, “executive functioning” refers to the ability to develop and sustain an organized, goal-directed, and flexible approach to problem situations. The executive functions allow individuals to adaptively use their basic skills (eg, core language skills, visual-perceptual ability, rote-memory capacity) in a complex and changing external environment. Operationalized here, this domain comprised tests demanding working memory, mental flexibility, planning, organization, behavioral inhibition, and problem solving. These data are consistent with those reported earlier by Engleman and colleagues. They are also consistent with a recent narrative literature review that asserted that executive functions are more vulnerable to disruption by OSA than are core intellectual or verbal abilities.

The Effect of OSA on Visual and Motor Skills was Inconsistent Across Analyses

The constructs of visual and motor ability are readily differentiated in theory, as the former places emphasis on cognitive processes, while the latter emphasizes simple motor execution. In practice, however, there are few tests of “pure” visual perception, and clinicians and researchers often rely upon tests of visual construction that require both cognitive processes and motor execution. This was reflected in our operational definition of the visual domain (Table 2). Given this, it is not surprising that effect sizes in the visual and motor domains behaved similarly. Within the control-referenced data set, tests of visual and motor ability displayed moderate to large effect sizes, ranging from .68 to 1.21. In contrast, the effect sizes were generally much smaller and non-significant in the norm-referenced data set.

Because the motor and visual domains also displayed considerable variability across studies, we posthoc explored the data for sources of this variability. This exploration suggested that OSA markedly affected fine-motor coordination and drawing but had much less effect on simple motor speed or visual perception. Though these exploratory findings remain to be prospectively tested, they do provide clues regarding the inconsistent findings across data sets. Within the motor domain, most of the case-controlled studies measured fine-motor coordination, whereas most uncontrolled studies measured simple motor speed. Within the visual domain, the 3 studies that included drawing tasks all included case controls.

The Effect of OSA on Memory Functioning was Inconsistent

The two data sets yielded discrepant memory findings. Whereas the control-referenced data set suggested moderate impairments in both short- and long-term visual memory (d = .56 and .55), the norm-referenced data set yielded small and nonsignificant effect sizes in both visual memory domains (d ≤ .14). Both data sets suggested that the impact of OSA on short-term verbal memory was small and not statistically significant (d ≤ .29). However, whereas the norm-referenced data set indicated moderately impaired long-term verbal memory (d = .52), the control-referenced data set yielded small and nonsignificant long-term verbal memory effects (d = .27).

It was difficult to make sense of the discrepant memory-test findings across data sets. Posthoc inspection of the data did not reveal any consistent associations between memory effect sizes and methodologic considerations (eg, test selection). However, it was clear that many studies used relatively old memory tests that have been sharply criticized for their psychometric shortfalls. Memory processes are complex— including input, organization, storage, retrieval, and output functions—yet few of the reviewed studies attempted to clarify exactly which aspects of memory are affected by OSA. It would be instructive for future researchers to employ recently developed memory tasks that allow for better differentiation of memory processes (eg, the California Verbal Learning Test). The recent paper by Salorio and colleagues represents a laudable attempt to untangle these processes. They reported that OSA-initiated executive-function deficits adversely impacted memory organization and retrieval but not long-term storage. They speculated that OSA may disrupt the integration of processes mediated by frontal and distal regions of the brain. This conclusion is similar to that reached by Harrison and Horne regarding the effects of sleep deprivation on memory functioning. However, in light of our mixed meta-analytic results, it will be important for future researchers to replicate and extend these findings within the OSA population.

Potential Moderators

We a priori identified 3 potential moderators of effect: publication status (peer-reviewed vs. non-peer-reviewed), study design (case-controlled vs. uncontrolled), and rough disease severity (mild or moderate vs. severe). None of these significantly moderated the mean norm-referenced effect sizes in the memory, executive, or motor domains. Study design moderated norm-referenced effect sizes in the intelligence and visual domains: OSA patients performed better in controlled studies than in uncontrolled studies. These moderator findings are difficult to interpret, however, because in all cases the mean effect sizes were in the neg-
The finding that disease severity did not moderate mean effect size across any domain is counter-intuitive in light of reports of an apparent dose-dependent relationship between OSA severity and neuropsychological outcome.\textsuperscript{5,9,30} Admittedly, the present approach of dichotomizing by mean disease severity in entire samples is crude and lacks sensitivity to effects across individuals. Moreover, our reliance upon the number of obstructive events as the sole index of severity was based upon the near-universal reporting of this information rather than on solid evidence that this is the best (or even the most consistently defined) index of OSA severity. Other factors, such as duration of events, degree of oxygen desaturation, or arousal frequency, may be more salient contributors to neuropsychological morbidity. Multivariate analyses of human data and controlled animal experiments will be necessary to clarify the disease features that are responsible for neuropsychological morbidity. Engelman and colleagues\textsuperscript{11} recently reviewed the available human data, noting that bivariate correlations between the severity of OSA patients’ sleep or respiratory disorder and their cognition, when reported, have generally been small to moderate and nested within a much larger matrix of non-significant correlations.

The relative lack of significant moderator findings does not imply homogeneity of effects across studies within a given domain. In fact, over half of the domains displayed variability across studies beyond what could be reasonably attributed to sampling variation. This both supports the use of random- and mixed-effects statistical models (rather than a fixed-effects model) and suggests the presence of other, unanticipated, determinants of effect size. We have speculated on some unanticipated variables above (eg, simple motor speed versus coordination) and suggest that differences in tests across studies substantially contributed to random-effects variance. This is partially a “methods” issue, as each test has unique psychometric qualities (eg, construct validity, reliability, distribution shape, sensitivity). However, this is also a conceptual issue, as each test assesses slightly different skills. For example, the executive-functioning domain, as defined here, may be broken down into multiple relatively homogeneous subdomains.\textsuperscript{4} We plan to examine these subdomains in greater detail in future work.

Implications for Mechanisms of Morbidity

The present data do not support a model of generalized neurologic dysfunction, as certain neuropsychological domains displayed clear impairment—vigilance, executive dysfunction, and motor coordination—while others displayed little impairment—intellectual, verbal, and perhaps visual-perceptual skills. Though the present data provide few clues as to specific etiologic pathways (eg, blood-gas abnormalities vs. sleep disruption), any model of etiology must account for this relatively specific pattern of neuropsychological morbidity. Beebe and Gozal\textsuperscript{4} recently proposed a model that could account for the executive dysfunction seen in OSA. Though this model focuses heavily upon the prefrontal cortex, the prefrontal region is richly connected with other cortical and subcortical structures, particularly subcortical gray matter, including the basal ganglia. Such interconnections have been implicated in other disorders that include both executive dysfunction and motor-coordination problems,\textsuperscript{3} and there is reason to believe that prefrontal connectivity is susceptible to sleep pathology.\textsuperscript{4} Another possible mechanism of morbidity is suggested by recent reports that the cerebellum plays a role not only in motor timing and coordination, but also in the coordination of higher-order cognitive processes.\textsuperscript{32,33} However, each of these models is in need of direct empirical study.

Implications for Clinical Practice

Like the medical sequelae of OSA (eg, cardiovascular complica-
tions\textsuperscript{34,35}), daytime neuropsychological sequelae should also be considered when making treatment decisions. In addition to diminishing immediate quality of life, the neuropsychological effects of OSA can have a long-term impact by the accumulation of scholastic, occupational, and relationship problems. Sleep clinicians should routinely inquire about vigilance, executive functioning, and motor coordination when working with patients presenting with signs of sleep-disordered breathing. Memory complaints may also reflect the effects of OSA, though it remains unclear what aspects of memory are most vulnerable. Because there is evidence that neuropsychological symptoms may persist despite sleep treatment,\textsuperscript{7,9} sleep clinicians should routinely assess these symptoms even following otherwise successful treatment. If residual deficits are suspected, this should lead to timely referrals and consultative involvement of mental health professionals. If a cognitive evaluation is undertaken, it should directly assess vigilance, executive functioning, and motor functioning, rather than rely on brief screening tools or tests of basic cognitive skills (eg, vocabulary) or intelligence.\textsuperscript{4} Conversely, individuals who display evidence of poor vigilance, motor-coordination problems, or executive dysfunction should be routinely screened for signs of sleep-disordered breathing (eg, loud persistent snoring, witnessed breathing pauses). Non–sleep-focused clinicians (eg, primary care physicians, psychologists) who treat adults who have displayed occupational, interpersonal, or cognitive declines concurrent with an increase in snoring or daytime sleepiness should refer patients to a sleep specialist who can conduct a thorough evaluation. When psychiatric or cognitive disorders are found to be comorbid with OSA, the sleep disorder deserves treatment in its own right.

Limitations

Any review, quantitative or qualitative, is reliant upon the quality of the research under consideration. Though there was not a strong a priori reason to question the general quality of research in this area, in the process of this review, we identified a number of factors that made it difficult to judge research quality. First, sample characteristics were often poorly detailed. Though demographic features such as race and education may have a negligible effect on many medical outcomes, they can have a profound effect on neuropsychological measures.\textsuperscript{15,36} Uncontrolled studies should more fully report sample demographics, and controlled studies should match for these variables. Second, disease characteristics of the samples were inconsistently described. Though the most relevant measures of disease severity remain unclear, we recommend that researchers routinely report variables identified as important in consensus statements (such as in practice parameters from the American Academy of Sleep Medicine\textsuperscript{37}), then supplement as they see fit. Third, research methods often lacked specificity, especially citations for the specific measures used. Psychological measurement is an evolving science, and multiple forms of any given measure may be available. The differences between forms can affect both norms and interpretation.\textsuperscript{19} Finally, reviews are hampered by incomplete reporting of results, especially the omission of raw data for effects that were not statistically significant. Important effects may be overlooked in small, underpowered studies that report only statistically significant results. Thus, while many of the reviewed articles were methodologically impressive, on other occasions it was difficult to assess research quality.

In addition to issues that can affect any review, the present review had several specific limitations. First, despite the large sample sizes in many analyses, moderator analyses were somewhat underpowered because the associated degrees of freedom were related to the number of samples (not subjects) at each level of the moderator. Moreover, as noted above, the only viable approach to studying the moderating effect of disease severity on outcome in this review was crude. Second, comparison against norms introduces additional sources of variance, as subjects in research studies often differ from normative groups in ways that may not be reflected by age, gender, or education. Potentially important differences may occur in the geographic origin of the norms, language, racial composition, and testing context. Thankfully, results across most
domains were relatively consistent across the control-referenced and norm-referenced data sets, increasing confidence that findings were fairly robust. Third, we were unable to perform norm-referenced analyses of the vigilance domain because of inadequate citations and the frequent use of nonstandardized or minimally normed measures. Fourth, there was inevitable confounding of domains with source samples, as no sample had data on all of the domains we reviewed. The use of a random-effects model accommodates this to some degree, resulting in broader confidence intervals than would be produced by a fixed-effects model. Nevertheless, comparisons of mean effect sizes across domains should be made cautiously and with due attention to the confidence intervals in Tables 3 and 4. Fifth, insofar as the domains under investigation were heterogeneous, true differences in effect size may occur across different tests of the domains. As noted above, this appeared to be the case in the motor domain and, perhaps, the visual domain as well. We are in the process of evaluating the multiple subcomponents of the heterogeneous construct of executive functioning.

Finally, the reader is reminded that the present findings were based upon group averages. Any given patient may display neuropsychological characteristics that differ from these averages because of the interaction of unique genetic, medical, environmental, and historical circumstances. The present findings provide a framework against which to compare individual patients, as well as general guidance for clinicians and researchers interested in the neuropsychological functioning of patients with OSA.

Concluding Comments

Despite these limitations, this review represents the most comprehensive overview of the neuropsychological effects of untreated OSA in adults published to date. Sophisticated mixed- and random-effects meta-analytic models minimized interpretive errors, allowed for investigation of potential moderators, and took into consideration a greater range of variability across studies than has been previously reported. Intelligence and basic verbal and visual-perceptual abilities were found to be resilient to the effects of OSA, whereas vigilance, executive functions, and motor coordination were found to be moderately to markedly affected. The effect of OSA on memory functioning was mixed. It is our hope that the findings inform the practice of clinicians who work with sleep-disordered patients and both stimulate and structure future research into the morbidity associated with OSA, including studies of the underappreciated effects of OSA on pediatric populations. We look forward to the products of that research.

ACKNOWLEDGEMENT

They authors gratefully acknowledge the helpful feedback of Dr. Michael Aloia, Dr. Cynthia Salorio, the Cincinnati Children’s Neuropsychology Team, and anonymous reviewers of an earlier draft of this manuscript.

REFERENCES

7. The Neuropsychological Effects of OSA—Beebe et al
44. Shipley WC. Institute of Living Scale. Los Angeles: Western Psychological Services; 1946.
APPENDIX

Studies Included in the Meta-Analysis

Bailey GL. Neuropsychological functioning, sleep and vigilance in men with obstructive sleep apnea syndrome treated with continuous positive airway pressure. [Dissertation]. Florida State University; 1993.


Morrier M. The relative influence of hypoxemia versus hypersomnolence on the behavioral morbidity of sleep apnea. [Dissertation]. Purdue University, West Lafayette, IN; 1991.


Walker CP. Neuropsychological changes in obstructive sleep apnea syndrome patients following nasal CPAP treatment. [Dissertation]. University of Alabama, Birmingham, AL; 1990.

NOTES

a) We did not discriminate “mild” from “moderate” obstructive sleep

...
apnea because only 3 clinical samples had an apnea-hypopnea index less than 15; all were community recruited.

b) Four studies reported at least some data in norm-referenced terms only (eg, norm-referenced z-scores) without accompanying raw scores, measure citations, or normative citations.38-41 In these cases, we adopted the original author’s norm-referenced effect sizes.

c) This is a conservative procedure for measures that are expected to cluster at one end of a distribution, including the poor scores often seen in clinical samples. Because measures within a psychological domain incompletely covary, a composite of multiple scores at the same end of a distribution is often more extreme than their simple arithmetic mean. There are more statistically precise ways of combining such scores,42 but this requires knowledge of covariance among the multiple measures within the samples, which was not presented in the source articles and which is difficult to infer from outside literature.

d) The assumption of better estimation with increased sample size may not apply when large samples are not representative of the population under scrutiny. As noted by others,4 this may have been the case with the study by Kim and colleagues,43 whose community-recruited sample with obstructive sleep apnea likely fell on the mild end of severity. This is a concern because their sample was of a magnitude several times larger than that of any other study and would, therefore, be quite influential in weighted mean-effect-size computations. Though random-effects and mixed-effects models tend to be more robust than fixed-effects models to such nonrepresentative samples, we supplemented the data reported in Tables 3 and 4 by computing unweighted mean effect sizes within the 4 domains affected by the Kim study: short- and long-term verbal memory, executive, and motor functioning. In the norm-referenced data set, the unweighted mean effect sizes were nearly identical to the weighted mean effect sizes reported in Table 3. In the control-referenced data set, the unweighted mean effect sizes ranged from .18 less to .27 more than the weighted mean effect sizes reported in Table 4. These control-referenced findings, which are consistent with those reported by Engleman and colleagues,11 do not substantially alter the conclusions of this study; vigilance, executive, and motor functioning displayed markedly greater effects than did any other domain.

e) For the domains of intelligence and long-term visual memory, too few mild-moderate OSA samples were represented to examine the impact of severity. Consequently, we reanalyzed all of the domains using an alternative cutoff for “severe obstructive sleep apnea” of an apnea-hypopnea index greater than 50 or an apnea index greater than 30. Though this resulted in better-equated samples at each level of severity, the impact of severity on outcome remained non-significant.