To evaluate the prevalence of mental disorders for persons of non-English-language origin, it is essential to use translated diagnostic interviews. The equivalence of translated surveys is rarely tested formally. In the National Latino and Asian American Study (NLAAS), the authors tested whether a carefully translated mental health survey administered in Spanish produced results equivalent to those obtained by the original English version, using a randomized survey experiment. The NLAAS is a nationally representative survey carried out in the United States in 2002–2003. Bilingual respondents from the Latino section of the NLAAS (n = 332) were randomly assigned to receive either a Spanish- or English-language version of the World Mental Health Survey Composite International Diagnostic Interview. In tests of differences in lifetime and 12-month prevalences of 11 diagnoses and four higher-order aggregate disorder categories, in only one case was there an apparent difference between randomized language groups: Lifetime reports of generalized anxiety disorder were more prevalent in the bilingual group assigned to English than in the group interviewed in Spanish. Detailed follow-up analyses did not implicate any specific question in the generalized anxiety disorder protocol. Translation and back-translation of surveys does not guarantee that response probabilities are exactly equivalent. Randomized survey experiments should be incorporated into cross-cultural psychiatric surveys when possible.

data collection; Hispanic Americans; language; mental health

Abbreviations: CIDI, Composite International Diagnostic Interview; DSM-IV, Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; GAD, generalized anxiety disorder; NLAAS, National Latino and Asian American Study; WMH, World Mental Health Survey.

Because substantial numbers, if not the majority, of persons afflicted with mental illness do not seek treatment (1–3), it is necessary to use surveys to determine the degree of psychiatric morbidity in the population rather than rely on clinical records. Participants in these surveys are asked to report their current and lifetime experiences with mental distress and psychiatric illness. This approach is especially useful when studying the health burden and needs of special
A number of strategies and methods have been developed to improve the quality of information obtained in face-to-face interviews and to connect this information to current diagnostic systems (4). Most of this methodological work has initially been done in English and subsequently tested with English-speaking populations. To study cross-cultural and cross-national variation in mental illness, researchers typically convene groups of experts to translate the English versions of surveys into languages varying from Spanish to Russian, Mandarin Chinese, and Yoruba (5). Guidelines for rigorous translation have been described (6, 7), and these guidelines attempt to take culture-specific expressions of distress and dysfunction into account when choosing words for the translated interview protocol.

There are many reasons why these well-intended procedures might not lead to assessment instruments that are comparable across cultures. Although the more rigorous procedures avoid verbatim translation, which can lead to awkward statements and ambiguous meanings, other, more subtle problems can arise when the culture and language of the respondent enters the picture (8). Even when translators agree on the best word for an English concept and when back-translators are able to recover the original word from the translation, it is still possible that respondents will interpret nuances or response options differently in the translated version. An example is the assessment of self-reported health, which does not seem to be as predictive of mortality risk for Latinos with low levels of acculturation as for Whites (9). Differences across languages in the perception of symptom severity or level of impairment could lead to changes in endorsement of illness criteria that consequently affect the overall prevalence of disorders. It is possible that some of the considerable variation in the cross-national rates of disorders recently reported by the World Health Organization (5) could be attributable to effects such as these.

In addition to the semantic issues related to translation, language can serve as an important cognitive priming cue that could make memories of culture-linked mental health experiences more accessible. These same cues might activate different representations of the respondents’ self-identities that are more or less consistent with the acknowledgment of mental disorders. Thus, the interview language represents an overall cognitive context that is culturally embedded, and this could influence responses, even if the translation equivalence is perfect.

Some of these issues have been considered in the context of evaluations of patients in clinical settings. In a classic set of papers, Marcos et al. (10, 11) compared severity ratings made by two English-speaking psychiatrists and two Spanish-speaking psychiatrists, who reviewed videographic recordings of standardized mental-status evaluations of 10 bilingual patients with probable schizophrenia. The patients had been interviewed both in Spanish and in English by separate psychiatrists in counterbalanced order. The investigators reported that ratings of the English segments of the interviews resulted in significantly greater severity than ratings of the Spanish segments. Although the studies by Marcos et al. raised important issues, the study design could not distinguish among effects of the patients’ language comprehension, the raters’ clinical interpretation of responses, and the role of the original clinical interviewers on the pattern of results.

In the context of epidemiologic surveys of mental disorders, there have been no formal tests of the impact of language on prevalence rates. In this article, we report the results of such a test of the Spanish translation of mental health measures used in the National Latino and Asian American Study (NLAAS) (12). Whereas randomized language studies have been applied to educational assessments (13), to our knowledge this is the first experimental evaluation of translation effects in a mental health survey.

In the NLAAS, investigators obtained a representative sample of Latino Americans living in the United States during 2002–2003 and determined that a portion of the study participants were able to respond to the survey in both English and Spanish. These bilingual participants were randomly assigned to receive either the English or the Spanish version of the interview. By comparing the prevalences of mental disorders reported by the randomly equivalent language groups, we were able to evaluate whether language influenced the reporting process and consequently the prevalence rates.

Our goals in this analysis were both to provide clear documentation of the comparability of the English and Spanish versions of the World Mental Health Survey (WMH) Composite International Diagnostic Interview (CIDI), as used in the NLAAS, and to illustrate how integrated survey experiments can be folded into epidemiologic studies of mental disorders to provide more precise evidence of cross-cultural comparability of results across different language groups. To give a complete picture of the possible effects of language on psychiatric surveys, we present results for both estimates of 12-month period prevalence and lifetime prevalence, although we acknowledge that interpretation of the latter quantity may be compromised by recall processes and confounding with the age distribution of any specific population (14).

MATERIALS AND METHODS

Sample

Details of the NLAAS sampling design are presented in the paper by Heeringa et al. (15) but are summarized here. In collaboration with the Institute for Social Research at the University of Michigan (Ann Arbor, Michigan), we obtained a representative sample of 2,554 persons in the United States who identified themselves as Latinos. The selection of a probability sample of Latino-American respondents followed a four-step sampling procedure designed to sample 1) US Census Metropolitan Statistical Areas and counties, 2) area segments, 3) housing units within the selected area segments, and 4) eligible respondents from the sampled housing units. The sample was selected and interviewed in 2002–2003.

Measures and translation

The full survey instrument has been described in detail by Alegría et al. (12). A central component of the instrument is
a version of the CIDI referred to as the WMH-CIDI (4). This instrument provides a structured assessment of a variety of Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (16), psychiatric diagnoses for the past 12 months and the respondent’s lifetime. The assessment makes use of general gate questions that inquire about certain kinds of distress and behaviors and then specific diagnostic questions that are asked of persons who answer the gate questions in the affirmative. In this report, we focus on 11 DSM-IV diagnoses: major depression, dysthymia, agoraphobia, panic disorder, generalized anxiety disorder (GAD), social phobia, post-traumatic stress disorder, alcohol dependence, alcohol abuse, drug (any substance) dependence, and drug (any substance) abuse.

The instrument also includes measures of functional impairment (the World Health Organization Disability Assessment Schedule (17, 18)) and utilization of mental-health services (the WMH-CIDI (4)). A variety of other measures relevant to characterizing the immigration experience and the experience of living as an ethnic minority in the United States are also included (19) but were not investigated for this report.

The WMH-CIDI was translated into Spanish and other languages in conjunction with a major cross-national study of mental health that was coordinated by the World Health Organization (5). For the NLAAS, the Spanish version of the core instruments was subjected to a further review following the principles described by Bravo et al. (6).

**Experimental design and procedure**

Interviewers employed in the NLAAS were certified to be fluent in both English and Spanish (20). They initiated the interview in the language used by the participant when he or she greeted the interviewer. Among the first questions asked was the degree to which the participant could speak Spanish and English. Adapting a format that had been previously used with Latino adolescents (21), participants were asked whether they spoke Spanish only, mostly Spanish (and some English), about the same amounts of Spanish and English, mostly English (and some Spanish), or English only. Respondents were then asked to select among these five language options. Participants who stated that they could not speak English or that they could speak only “some English” were classified as monolingual Spanish speakers. They were administered the interview in Spanish. Likewise, those who could not speak Spanish or could speak only “some Spanish” were classified as monolingual English and were administered the interview in English. Those who reported speaking both Spanish and English “about the same” were classified as bilingual respondents. The interviewer’s computer was programmed to randomly assign bilingual interviewees to receive either the Spanish version or the English version, which we refer to as the “Spanish arm” and the “English arm” of the study. Interviewers were instructed to maintain the interview exclusively in the randomized language and were discouraged from switching to the other language, unless the interviewee indicated that he or she was not bilingual or wanted to terminate the interview. This did not happen in any case, so the interviews for the bilingual Latinos were conducted in the assigned language and were not subject to toggling to the alternative language.

**RESULTS**

**Characteristics of language groups**

Among the 2,554 Latino participants in the NLAAS, 1,348 (52.8 percent) represented themselves as monolingual Spanish speakers, 874 (34.2 percent) represented themselves as monolingual English speakers, and 332 (13.0 percent) represented themselves as bilingual. Of the last group, 150 were randomly assigned to the Spanish arm of the study and 182 to the English arm. The difference in the sizes of these groups was consistent with binomial variation (Pearson’s $\chi^2 (1 \text{ df}) = 3.08; p > 0.07$).

Table 1 shows the demographic characteristics of the two groups, as well as $\chi^2$ test statistics for the hypothesis that the two distributions were the same. As expected, the groups were equivalent with regard to the majority of comparisons in table 1. However, two comparisons were significant at the 0.05 level. There were somewhat fewer US-born bilingual respondents assigned to the Spanish arm than to the English arm (40.7 percent vs. 51.7 percent; $\chi^2 (1 \text{ df}) = 3.98; p = 0.046$). We also found some evidence that participants in the English arm reported better English proficiency than those in the Spanish arm ($\chi^2 (3 \text{ df}) = 9.52, p = 0.023$). In contrast to the imbalance in nativity, this difference could have been due to the experience of the interview. Both English and Spanish proficiency were determined by self-report of expertise with the language; these reports were obtained at the end of the interview and hence could have been influenced by the interview experience. Because of the trends for imbalance of the groups with regard to nativity and English language proficiency, we adjusted subsequent comparisons for these variables.

**The effect of language on prevalence estimates**

Table 2 shows the 12-month and lifetime prevalence rates of DSM-IV disorders among bilingual respondents according to the Spanish and English versions of the WMH-CIDI, with adjustment for differences in nativity and English language proficiency. Across the majority of diagnoses, both 12-month and lifetime, the Spanish and English versions of the WMH-CIDI yielded similar rates among the bilingual participants who were randomly assigned to one language or the other. For example, the 12-month period prevalence of any disorder was 18.0 percent among persons answering in Spanish and 17.0 percent among persons answering in English. The proportion of the sample who reported experiencing an episode of any of the disorders during their lifetime was 30.2 percent among Spanish bilinguals versus 30.8 percent among English bilinguals.

A possible exception to this trend was reported lifetime experience of GAD. Whereas 8.2 percent of the bilingual respondents who answered in English reported a lifetime diagnosis of GAD, only 4.2 percent of those who answered
in Spanish reported this history (odds ratio = 0.38, 95 percent confidence interval: 0.13, 1.07; p = 0.07). For 12-month GAD, the 95 percent confidence interval for the odds ratio (95 percent confidence interval: 0.02, 1.13) was consistent with the difference in lifetime rates but was also consistent with equal rates. We checked to see whether the difference was due to any of the four CIDI gate questions that asked about lifetime history of 1) being a worrier, 2) being unusually nervous or anxious, 3) experiencing a period in which one was anxious/worried on most days, and 4) having excessive worries or nervousness. None of these questions showed an imbalance across languages. Next we looked at the specific DSM-IV criteria for GAD (16), which involves six components. All of these six components were fulfilled more in the English language group than the Spanish language group, which means that none of them had a translation problem that could uniquely account for the GAD language difference. The largest difference was observed for questions about excessive anxiety and worry, anxiety occurring on more days than not for at least 6 months, and anxiety involving a number of events or activities. These questions contributed to what the DSM-IV (16) calls GAD criterion A, which was met by 9.3 percent of the respondents in the Spanish arm and 13.2 percent of those in the English arm. This criterion involved a time element—"the longest period of months or years in a row."
To see whether the language effects were apparent in the results among NLAAS respondents who were not randomized to a language arm, we examined differences in GAD among monolingual Spanish respondents and monolingual English respondents. The pattern did not carry over to these groups. The lifetime rate of GAD was 6.2 percent (standard error, 0.6) for the former group and 5.7 percent (standard error, 0.7) for the latter group.

**DISCUSSION**

The literature on cross-cultural measurement of mental disorders has focused on the important issues of literate translations of English questions that do justice to idiomatic expressions within specific language communities (22, 23) and the internal psychometric features of the translated instruments (24), such as test-retest reliability. Although these are important steps for the establishment of cross-cultural comparability of measures, they do not guarantee that individuals will provide the same responses to questions posed in the new language as they would in the first language (25). When bilingual respondents are available in a study population, it is possible to directly test the equivalence of measures by employing randomized survey designs (26, 27).

To our knowledge, our study was the first to use a randomized survey experiment to test whether a carefully prepared

---

**TABLE 1. Continued**

<table>
<thead>
<tr>
<th></th>
<th>Monolingual Spanish speakers</th>
<th>Bilingual Spanish speakers</th>
<th>Bilingual English speakers</th>
<th>Monolingual English speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of persons</td>
<td>%</td>
<td>No. of persons</td>
<td>%</td>
</tr>
<tr>
<td>Spanish speaking</td>
<td>24</td>
<td>1.8</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Poor</td>
<td>231</td>
<td>17.2</td>
<td>21</td>
<td>14.1</td>
</tr>
<tr>
<td>Fair</td>
<td>550</td>
<td>40.9</td>
<td>64</td>
<td>43.0</td>
</tr>
<tr>
<td>Good</td>
<td>540</td>
<td>40.2</td>
<td>64</td>
<td>43.0</td>
</tr>
<tr>
<td>Excellent</td>
<td>58</td>
<td>4.3</td>
<td>7</td>
<td>4.7</td>
</tr>
<tr>
<td>Spanish reading</td>
<td>215</td>
<td>16.0</td>
<td>29</td>
<td>19.5</td>
</tr>
<tr>
<td>Poor</td>
<td>554</td>
<td>41.2</td>
<td>52</td>
<td>34.9</td>
</tr>
<tr>
<td>Fair</td>
<td>518</td>
<td>38.5</td>
<td>61</td>
<td>40.9</td>
</tr>
<tr>
<td>Good</td>
<td>91</td>
<td>6.8</td>
<td>12</td>
<td>8.1</td>
</tr>
<tr>
<td>Excellent</td>
<td>58</td>
<td>4.3</td>
<td>7</td>
<td>4.7</td>
</tr>
<tr>
<td>Spanish writing</td>
<td>264</td>
<td>19.6</td>
<td>36</td>
<td>24.2</td>
</tr>
<tr>
<td>Poor</td>
<td>550</td>
<td>40.9</td>
<td>49</td>
<td>32.9</td>
</tr>
<tr>
<td>Fair</td>
<td>440</td>
<td>32.7</td>
<td>52</td>
<td>34.9</td>
</tr>
<tr>
<td>Good</td>
<td>91</td>
<td>6.8</td>
<td>12</td>
<td>8.1</td>
</tr>
<tr>
<td>Excellent</td>
<td>58</td>
<td>4.3</td>
<td>7</td>
<td>4.7</td>
</tr>
<tr>
<td>English speaking</td>
<td>796</td>
<td>59.3</td>
<td>3</td>
<td>2.0</td>
</tr>
<tr>
<td>Poor</td>
<td>352</td>
<td>26.2</td>
<td>18</td>
<td>12.1</td>
</tr>
<tr>
<td>Fair</td>
<td>141</td>
<td>10.5</td>
<td>73</td>
<td>49.0</td>
</tr>
<tr>
<td>Good</td>
<td>53</td>
<td>4.0</td>
<td>55</td>
<td>36.9</td>
</tr>
<tr>
<td>Excellent</td>
<td>796</td>
<td>59.3</td>
<td>4</td>
<td>2.7</td>
</tr>
<tr>
<td>English reading</td>
<td>342</td>
<td>25.5</td>
<td>15</td>
<td>10.1</td>
</tr>
<tr>
<td>Poor</td>
<td>148</td>
<td>11.0</td>
<td>67</td>
<td>45.0</td>
</tr>
<tr>
<td>Fair</td>
<td>56</td>
<td>4.2</td>
<td>63</td>
<td>42.3</td>
</tr>
<tr>
<td>Good</td>
<td>872</td>
<td>65.0</td>
<td>6</td>
<td>4.0</td>
</tr>
<tr>
<td>Excellent</td>
<td>303</td>
<td>22.6</td>
<td>20</td>
<td>13.4</td>
</tr>
<tr>
<td>Good</td>
<td>17</td>
<td>11.0</td>
<td>67</td>
<td>45.0</td>
</tr>
<tr>
<td>Excellent</td>
<td>50</td>
<td>3.7</td>
<td>56</td>
<td>37.6</td>
</tr>
<tr>
<td>Total</td>
<td>1,348</td>
<td>150</td>
<td>182</td>
<td>874</td>
</tr>
</tbody>
</table>
translated version of an epidemiologic battery was empirically equivalent to the version prepared in the original language (English). This experiment helps address the question of whether literate translated phrasings of questions evoke the same or different recollections of symptoms of psychiatric illnesses as the original questions. This is a preliminary step toward determining whether the measure distinguishes between psychiatric illnesses and culturally determined responses to illness.

The results for the Spanish and English versions of the WMH-CIDI were comparable for 10 of the 11 diagnoses considered here. For these diagnoses, there was no evidence...
that the rates in the Spanish arm differed from those obtained in the English arm for bilingual respondents.

The one possible exception was GAD, which seemed to have a higher lifetime rate among those bilingual respondents assigned to the English arm. We looked for evidence that the translation of a specific question or set of questions could account for this difference, and we found none. The excess rate in the English-version group was spread across questions and decision rules. We also checked to see whether this excess was consistent with differences between monolingual English respondents and monolingual Spanish respondents, and we found that the pattern was not consistent in these comparisons. In fact, monolingual English respondents had slightly lower rather than higher rates of GAD. When all four groups are compared together, it seems that the bilingual respondents assigned to the English arm are the ones whose lifetime rate (8.2 percent) is unusual in comparison with monolingual English (5.7 percent), bilingual Spanish (4.0 percent), or monolingual Spanish (6.2 percent) respondents. Because the risk of Type I error increases with multiple significance tests, and given the fact that we considered 30 comparisons in table 2, we cannot rule out the possibility that the difference in GAD is simply due to sampling fluctuation.

There were several limitations to our approach. One is that our comparisons focused only on the concrete response probabilities and the use of this information to generate prevalence rates. Even when these rates are strictly comparable across language groups, reasonable questions can still be raised about the extent to which Western Anglo-American psychiatric nosology can be applied universally across the globe.

Another limitation is that we cannot guarantee that the response processes in the bilingual group were accurate reflections of what happens with monolingual groups. Bilingual respondents tend to be more educated and more complex in their cultural identity, and hence they might have had unique responses to the questions we asked them. There is an active research literature in linguistics that aims to determine how language processing is affected by knowledge of more than one language. Subsequent studies of the psychometric item response patterns in the four language groups could help determine whether there was any evidence of differential item functioning across the groups. However, the psychometric methods make assumptions about how response patterns can be modeled that are not needed in the experimental analysis presented here.

A final limitation, an important one, is that our results cannot establish that the various diagnostic algorithms are uniformly equivalent across the two language versions. Our method made use of available bilingual respondents in a survey of Latino Americans and was not designed to make a definitive statement about the absolute equivalence of the translations. We presented confidence intervals for the contrasts between the language versions to provide readers with information about the degree of certainty that the versions would produce the same prevalence results. Although the majority of the 95 percent confidence intervals in table 2 included a value consistent with the versions’ being equivalent, a number of the intervals also included values that would be consistent with important differences. This occurred for extremely rare disorders for which we had little power to detect most differences. For a number of other disorders, however, the 95 percent confidence intervals suggest that our data are inconsistent with large differences (e.g., odds ratios of 3 or more.) The power and precision of our analyses were affected by the relatively modest number of bilingual respondents who were identified in our representative sample.

Even with these limitations, we believe that the use of language experiments in health surveys of multilingual populations provides a feasible step toward testing the translation equivalence of the survey instruments. Routine tests of translational equivalence of major surveys are consistent with good management goals of continuous quality improvement of products and services. The feasibility of continuous quality improvement procedures such as survey experiments is increased by the widespread use of electronic data collection methods, such as laptop computers, which can be programmed to implement rigorous randomization designs. As epidemiologists and health policy analysts turn toward the problem of documenting risk differences and reducing service disparities that affect immigrant groups, it is essential to adopt routine checks on the equivalence of reports obtained in different languages to rule out spurious influences on patterns of findings.

Perhaps most critically, the finding of scant translation effects from this experiment adds confidence in the validity of the comparisons of NLAAS groups that differ with regard to the proportion of persons who provided information in Spanish versus English. Recent reports have been able to document important differences in the mental health morbidity of Latinos of different national origins and to provide a clearer description of the so-called immigrant paradox, whereby economically disadvantaged immigrants appear to have unusually good health relative to Anglos and children of immigrants. Both basic epidemiologic inferences and health-services analyses benefit from rigorous tests of language effects.

ACKNOWLEDGMENTS

This project was supported by National Institutes of Health research grants U01 MH062209 and P50 MH073469, both funded by the National Institute of Mental Health.

The NLAAS data used in this analysis were provided by the Center for Multicultural Mental Health Research of the Cambridge Health Alliance (Cambridge, Massachusetts).
Conflict of interest: none declared.

REFERENCES