Contra Costa County Women Infants & Children (WIC)
Perinatal Depression Screening, Education and Referral Project

WIC PHQ9 and PHQ4 Validation Study Report

Elaine Zahnd, PhD, Independent Consultant
February 6, 2011

This validation study report consists of two components:

I. *Conduct a literature review on any validation studies that have been completed on the PHQ-4 to determine how the populations included in the studies compare to the WIC female low-income population.*

II. *Conduct a small validation study with WIC clients comparing outcomes for the PHQ-9 and PHQ-4 to see if consistent findings for the WIC population. WIC staff ask clients to complete PHQ-9 and PHQ-4 at the same visit (randomize which comes first), and score results for a minimum of 50 clients in English and Spanish. Analyze results to determine risk levels using the new shorter screener and how it compares to the longer PHQ-9 screener. Summarize results and make recommendations about usefulness of the PHQ-4 for the WIC population.*

I. Literature review on existing validation studies completed on the PHQ-4 to determine how populations included in such studies compare to Contra Costa County WIC population (female, racially/ethnically diverse, pregnant and parenting, low-income).

Background on the Patient Health Questionnaires (PHQ-9, PHQ-2, PHQ-4)

In the mid-1990s, Robert L. Spitzer, MD, Janet B.W. Williams, DSW and Kurt Kroenke, MD, along with Columbia University colleagues, created the PRIME-MD or "Primary care evaluation of mental disorders diagnostic tool". It had modules for 12 different mental health disorders. While undergoing work on the PRIME-MD, they developed the Personal Health Questionnaire (PHQ) and the Generalized Anxiety Disorder (GAD-7) screeners for use with the general population. The PHQ-9 is a screening tool focused specifically on detecting depression, while the GAD-7 measures 7 common anxiety symptoms. From these initial screeners, the PHQ-2 and the GAD-2 (each a two item measure) were developed respectively to screen quickly and accurately for depression and general anxiety. The need was to develop very short screeners to

encourage primary care providers and other health care and social services staff to utilize them in every day practice settings.

The PHQ-4 was developed as an ultra-brief screener for anxiety and depression for use during outpatient or home visits. It was also tested with obstetric-gynecology populations to see how well it performed with women during pregnancy and up to one year postpartum. It can be administered by a doctor, other health care personnel or can be self-administered. It combines two validated screeners (see PHQ section below). The PHQ-4 is a screening tool; it is meant to alert providers to the need for further assessment and diagnosis and/or referrals. The PHQ-4 does not diagnose depression or anxiety.

Targeting pregnant women and mothers of young children for depression screening, especially during the postpartum period, has been a primary depression screening focus according to a recent report from the National Academy of Sciences, especially since the emergence of such depression can impact both parenting and child development. (National Research Council and Institute of Medicine of the National Academy of Sciences, Depression in Parents, Parenting, and Children: Opportunities to Improve Identification, Treatment and Prevention, 2010, www.nap.edu/catalog/12565.html).

Most adults with depression do not get treated for it. Kessler's large epidemiological study revealed that less than 1/3 of adults with major depression have accessed general medical or specialty emotional or mental health outpatient services in the previous year (Kessler, RC, Zhao, S, Katz, SJ, Kouzis, AC, Frank RG, Edlund, M, and Leaf, P. Past-year use of outpatient services for psychiatric problems in the National Comorbidity Survey, American Journal of Psychiatry, 1999, 156: 115-123). The one-third of adults with major depression who do get treated use a variety of alternative points of contact and entry, noteworthy given the Contra Costa County WIC Perinatal Depression Project (Kessler RC, Merikangas KR, and Wang PS. Prevalence, comorbidity and service utilization for mood disorders in the United States at the beginning of the twenty-first century, Annual Review of Clinical Psychology, 2007, 3: 137-158).

According to the literature, when screening parents for depression, screening tools are often used without validation by other methods. Kroneke and colleagues (see PHQ section below) found that screening tools detect approximately 75% of the general population with any type of depression and 25-40% of those with a major depressive disorder. Of note, the use of a screening tool by itself to define depression within a particular population, such as a pregnant or parenting WIC population, generally underestimates the prevalence of depression; further assessment and diagnosis is needed. Relying on symptoms alone, however, may do the opposite – basically, overestimate the prevalence, especially since symptoms wax and wane during the perinatal period.

Common tools for screening for depression in a perinatal population include the CES-D or Center for Epidemiologic Studies Depression Scale, the Edinburgh Postnatal Depression Scale (EPDS), the Beck Depression Inventory, and the Patient Health Questionnaire 2 (PHQ-2). These screeners vary in their ease of administration in clinical, primary care or other social service settings. Some are considerably longer screeners than others. Although the Edinburgh scale has been widely promoted for screening during the perinatal time frame, according to Gaynes and colleagues, it does not have good sensitivity for detecting major or minor depression among the perinatal population. They also noted that previous studies on the Edinburgh scale have shown varying cut point scores, an issue that can prove confusing. Of greater importance in judging its usefulness according to Gaynes and colleagues, however, is that it is rarely is validated with a clinical diagnostic interview (Gaynes, BN, Gavin, N, Meltzer-Brody S, Lohr, KN, Swinson, T, Gartlehner, G, Brody, S, and Miller, WC, Perinatal Depression: Prevalence, Screening Accuracy, and Screening Outcomes [Evidence Report/Technology Assessment # 119, AHRQ publication # 05-E006-2], Rockville, MD: Agency for Healthcare Research and Quality, 2005). In fairness, newer screening tools are often validated and then put into wide spread practice without further validation with specific populations due to cost, difficulty of obtaining funding for methodological studies, and the time involved in setting up a validation study using clinical diagnostic interviews or other outcome validation measures.

The discussion above is meant to emphasize the fact that it is difficult to measure the magnitude and severity of postpartum depression and the emotional and mental health needs of mothers postpartum *without further assessment beyond the screening tools under scrutiny.* Despite these omissions, screeners are an important part of the overall provider toolbox of detection, assessment, diagnosis, referrals, further assessment of possible comorbid conditions, and the culmination in the provision of needed services.

The lack of available and effective resources for those suffering from depression is one of the major concerns when a perinatal depression screening program is undertaken, especially given the related risk factors of maternal depression on the family and children in the home. The earlier evaluation report assessing how well the PHQ-9 performed with the Contra Costa County WIC population included a staff evaluation piece of the pilot project, and a number of staff noted their concern about the need for developing more accessible and effective resources for the women screened using the PHQ-9 (Contra Costa County WIC Perinatal Depression Screening, Education and Referral Project, Project Evaluation Final Report, Zahnd, Elaine, 10-7-10).

In attempting to determine the degree to which physicians screen for maternal or perinatal depression overall, it appears that the use of written depression screening tools is still fairly rare. Instead many doctors report that they use observation or brief inquiries instead of

written tools during patient visits; this is especially true during child wellness visits (National Research Council and Institute of Medicine of the National Academy of Sciences, Depression in Parents, Parenting, and Children: Opportunities to Improve Identification, Treatment and Prevention, 2010, www.nap.edu/catalog/12565.html).

Patient Health Questionnaire (PHQ) Validation Studies

The literature on screening for postpartum depression has emerged largely from population studies rather than clinical trials. An early meta-analysis of screening studies shows an average rate of about 13% for postpartum depression overall (O'Hara and Swain, 1996). The rate is higher for women with a history of depression. Factors increasing risk for perinatal depression include being a single parent (of which the WIC population has a fair percentage), unplanned pregnancies, lack of social support and poor financial resources (National Academy of Sciences, ibid, p. 189). A recent PRAMS CDC study in 17 states showed a range of self-reported postpartum depression from 12 to 20% (Pregnancy Risk Assessment Monitoring System Working Group and the Centers for Disease Control and Prevention Pregnancy Risk Assessment Monitoring System Team, 2008). Gaynes et al. meta-analysis found 30 studies providing mainly point prevalence rates of perinatal depression; the range for *major depression* during the perinatal period was from 3.1% to 4.9% during various pregnancy periods, and from 1.0% to 5.9% at various times during the 1st postpartum year. For minor and major depression combined the range was 8.5%-11% during different times during pregnancy and from 6.5% to 12.9% during varying times for the 1st postpartum year. Confidence intervals were wide, so there is still a great deal of uncertainty (Gaynes, BN, Gavin, N, Meltzer-Brody S, Lohr, KN, Swinson, T, Gartlehner, G, Brody, S, and Miller, WC, Perinatal Depression: Prevalence, Screening Accuracy, and Screening Outcomes [Evidence Report/Technology Assessment # 119, AHRQ publication # 05-E006-2], Rockville, MD: Agency for Healthcare Research and Quality, 2005).

The carefully designed Gaynes et al. meta-analysis also posed the question: *What is the accuracy of different screening tools for detecting depression during pregnancy and the postpartum period?* However, no versions of the PHQ were included, mainly because they conducted a systematic review with a high quality level of inclusion and exclusion criteria, and thus out of an original search identifying 846 studies, only 109 were pulled for final review, and of those on 23 studies remained in the review for this issue. They focused on sensitivity and specificity of the screening tools (sensitivity = proportion of patients with a condition/disease (in this case, depression) who test positive or "true positives"; specificity = proportion of patients without the condition/disease who test negative or "true negatives". However, when they addressed the question in detail, they reported only 10 studies that reported test characteristics of English-language screeners. These 10 had fair to good quality but external

validity, a recurrent problem, was poor to fair. The study populations were almost entirely white, a major limitation. There were also few depressed patients that made it impossible for them to identify ideal cut off points. One study using the Edinburgh Postnatal Depression Scale (EPDS) only had 6 patients with major depression, and 14 with either major or minor depression. Specificity was good (.72 to .95) but sensitivity was poor (.57 to .71). As the authors noted:

For postpartum depression, also, the small number of depressed patients involved in the studies precluded identifying an optimum screener or optimum threshold for screening (Gaynes et al., Ibid. p. 4).

In conclusion, the authors note that the imprecision among all the study screening instruments they reviewed for major depression (Beck Depression Inventory [BDI], Postpartum Depression Screening Scale [PDSS], and Edinburgh [EPDS] meant they could not say the sensitivity estimates among the different tools were different. They noted that providers need to determine if falsely missing depression is worse than falsely identifying it. The findings for major or minor depression showed high specificity but low sensitivity, so much so that they could not determine if one screener or cutoff performed differently than any of the others. With only 15 studies for addressing whether screening leads to improved patient outcomes, of which only 11 were judged of fair quality, their main concern was the lack of racial/ethnic diversity in the samples and the lack of power to demonstrate statistically significant differences. Basically, the research on screener performance and on improved outcomes is still very much in its infancy.

Turning to the PHQ-4 validation studies specifically, one article should be highlighted. This article describes the initial validation study of the newly-constructed PHQ-4 (Kroenke MD, Kurt, Spitzer MD, Robert L, Williams DSW, Janet BW, and Lowe MD, PhD, Bernd, An Ultra-Brief Screening Scale for Anxiety and Depression: The PHQ-4, Psychosomatics, 50:6:613-621, Nov-Dec 2009). The authors sought to test an extremely brief screener for utilization in clinical and primary care settings. They combined two validated brief screeners for anxiety and depression drawing from the Patient Health Questionnaire consisting of 9 questions (PHQ-9), and the Generalized Anxiety Disorder 7 scale (GAD-7). From the PHQ-9, they used the PHQ-2, a two-item measure that includes the core criteria for depression. Drawing from the GAD-7, they utilized the shorter two item measure for anxiety (GAD-2). Both the PHQ-2 and the GAD-2 have been shown to be excellent screeners, using a cut point of 3 or greater on the PHQ-2 scale (sensitivity of 83%; specificity 90%) for major depression disorders, and the same cut point on the GAD-2 (88% sensitivity for generalized anxiety disorder; specificity 81-83% for four disorders, specifically generalized anxiety, panic, social anxiety, and posttraumatic stress disorders). The new 4 item (PHQ-4) was compared with longer depression and anxiety

measures to determine construct validity; they used the Medical Outcomes Study short-form general health survey (SF-20), a measure of health functioning as well using number of doctor visits and disability days during the previous 3 months. They also drew a random sample of 965 subjects and had health professionals conduct blinded DSM-IV telephone interviews to determine independent diagnoses. The results confirmed that depression and anxiety explained 84% of the total variance, and that an increase in PHQ-4 scores was strongly associated with functional impairment, and days of disability or days of healthcare. They concluded that the PHQ-4 was a valid screening instrument for detecting both anxiety and depression disorders.

Below is a presentation of the sample population for the PHQ-4 validation study and how it compares with the WIC population:

Sample Population: 2149 patients from 15 primary care sites in 12 different states in the U.S. The 15 primary care sites included 13 family practice sites and two internal medicine sites. Mean age 47.2 years (range 18-95); 66% female, 81% non-Hispanic white, 8% African American, 8% Hispanic; 7% < high school; 31% HS graduates; 62% some college; 64% married.

WIC Population for Contra Costa County Depression Project: Differences include gender (WIC 100% female), race/ethnicity (WIC sample is more diverse with larger populations of Hispanic and African American females), pregnant or with a child under 5 years or an infant under 12 months (a younger population), low income (below 185% of FPL guidelines).

Of note, since this initial validation study on the PHQ-4 was designed to demonstrate its validity and utility with all primary care populations, it is understandable that it would not focus on a population like WIC. Earlier work in the development of the PHQ-2 and PHQ-4 utilized a large sample of patients in primary care clinics (C. Spitzer RL, Kronenke K, Williams JBW. Patient Health Questionnaire Study Group. Validity and Utility of a self-report version of PRIME-MED Study. JAMA 1999; 283:1737-44 [PubMed])

Sample Population: 3000 patients in 8 primary care clinics.

(D. Spitzer, RL, Williams JBW, Kroenke K, et al. Validity and utility of the Patient Health Questionnaire in assessment of 3000 obstetric-gynecology PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. Am J Obstet Gynecol. 2000; 183:759-69 [PubMed])

Most investigation of the PHQ has focused on the PHQ-9; the PHQ-4 is relatively new in comparison, however, the work validating the PHQ-9 is of relevance. The following study is widely regarded as the main validation study on the PHQ-9 demonstrating the reliability, construct and procedural validity for depression disorders. (Kroneke K, Spitzer RL, Williams

JDW: The PHQ 9: validity of a brief depression severity measure. J Gen Intern Med 2001; 16:606-613). The sample population included a large number of patients from seven obestric-gynecological health clinics, providing some solid support of the usefulness of the PHQ-9 with women of reproductive ages; however it was not specifically focused on a perinatal population. The population drawn from the ob-gyn clinic sites more closely mirrored the WIC population than other studies in that besides being entirely women, they were younger group of reproductive age, more were Latina or Hispanic, they had a lower average education and less medical comorbidity compared to the primary care clinic sample. In addition, before seeing the physician, all patients completed the SF-20 and the PHQ and estimated the number of doctor visits and disability days they had experienced in the previous 3 months. The internal reliability of the PHQ-9 was excellent (Cronbach alpha of 0.89) and test-retest reliability was also excellent. This validation study had a relatively young, disproportionately female sample. Data from the two related studies with 6000 patients provided strong evidence for the validity of the PHQ-9 as a brief depression screener. They also noted the need for more longitudinal studies since this was cross-sectional.

Sample Population: 6000 patients in 8 primary care clinics and 7 obstetrics-gynecology clinics.

In summary, as shown by the lack of validation studies of the PHQ-4 with perinatal populations, there is a need for further validation studies. Recommendations concerning the usefulness of the screener will follow the 2nd component of this report.

II. Conduct a small validation study with WIC clients comparing outcomes for the PHQ-9 and PHQ-4 to see if consistent findings for the WIC population. WIC staff ask clients to complete PHQ-9 and PHQ-4 at the same visit (randomize which comes first), and score results for a minimum of 50 clients in English and Spanish. Analyze results to determine risk levels using the new shorter screener and how it compares to the longer PHQ-9 screener. Summarize results and make recommendations about usefulness of the PHQ-4 for the WIC population.

In late November, an effort began to administer the PHQ-4, a shorter screener than the PHQ-9 at WIC sites countywide. The PHQ-9 had been used at WIC sites since the pilot project, "Contra Costa County Women Infants & Children (WIC) Perinatal Depression Screening, Education and Referral Project", began in May of 2010. The purpose of the PHQ-4 project was to determine how well the shorter screener worked in comparison to the PHQ-9. If the PHQ-4 had comparable results, it meant that WIC staff could consider the adoption of the shorter PHQ depression screener. This project collected data on the PHQ-9 and the PHQ-4 risk level results

for the WIC population at all four WIC sites: Richmond, Pittsburg, Brentwood and Concord. The PHQ-4 combines two questions from the PHQ-9 on depression, and two questions from the GAD-7 on anxiety. Specifically, for the PHQ-4, each WIC client is asked:

Over the last 2 weeks, how often have you been bothered by the following problems?

1. *Feeling nervous, anxious or on edge (anxiety)*
2. *Not being able to stop or control worrying (anxiety)*
3. *Little interest or pleasure in doing things (depression)*
4. *Feeling down, depressed or hopeless (depression)*

While the design called for a small comparison study of 50 clients who were administered both the PHQ-4 and PHQ-9, the WIC staff administered the PHQ-4 to a total of 1379 WIC clients between December 1, 2010 and December 20, 2010.

At the same time that the WIC client filled out a PHQ-4 screener, they also completed a PHQ-9 screener. Although the PHQ-9 has five rather than the four risk levels that exist for the PHQ-4, two similar PHQ-9 risk levels, specifically "moderately severe depression" (15-19) and "severe depression" (20-27) were combined in order to match the risk level categories for the PHQ-4 for this comparative analysis. Scoring for the two screener risk levels is shown on Table 1 below:

TABLE 1
SCORING FOR PHQ-9 AND PHQ-4

Risk Level	PHQ-9 Scoring	PHQ-4 Scoring
None-Minimal	0-4	0-2
Mild	5-9	3-5
Moderate	10-14	6-8
Severe	15-27	9-12

Table 2 through Table 5 findings are presented below to illustrate the outcomes per site for the four risk levels of the PHQ-4 for all four county WIC sites:

TABLE 2

BRENTWOOD	PHQ-4 Depressive Severity Score				
Scoring	0-2	3-5	6-8	9-12	TOTAL
Total WIC Clients	174	22	7	6	209
Percentage	83%	11%	3%	3%	100%

TABLE 3					
CONCORD	**PHQ-4 Depressive Severity Score**				
Scoring	**0-2**	**3-5**	**6-8**	**9-12**	**TOTAL**
Total WIC Clients	315	35	19	7	376
Percentage	84%	9%	5%	2%	100%

TABLE 4					
PITTSBURG	**PHQ-4 Depressive Severity Score**				
Scoring	**0-2**	**3-5**	**6-8**	**9-12**	**TOTAL**
Total WIC Clients	421	57	27	8	513
Percentage	82%	11%	5%	2%	100%

TABLE 5					
RICHMOND	**PHQ-4 Depressive Severity Score**				
Scoring	**0-2**	**3-5**	**6-8**	**9-12**	**TOTAL**
Total WIC Clients	248	36	10	7	301
Percentage	82%	12%	3%	2%	100%

As shown in the four site-based tables above, there were 209 WIC clients at Brentwood, 376 WIC clients at Concord, 513 WIC clients at Pittsburg, and 301 WIC clients at Richmond who were administered the PHQ-4 during the month of December 2010.

There are only small differences in the range of scores by risk category by site, indicating that the instrument performs equally well across county sites despite being administered at different geographic sites with somewhat different populations. Thus, the findings demonstrate good comparability across the sites despite the different racial/ethnic and other demographic characteristics that exist for each of the four regions. As Tables 2-5 show, the range for "no or minimal risk" of depressive symptoms is from 82% to 84%. Similarly, "mild risk" ranges from only 9% to 12% among the sites, "moderate risk" ranges from 3% to 5%, and "severe risk" ranges from 2% to 3%. These results indicate that the PHQ-4 does not vary by regional administration.

As shown in Table 6 below, the results for the WIC-administered PHQ-4 reveals that of a total of 1399 WIC clients, 83% of the WIC clients had "no or minimal" risk of depression, 11% screened in at the "mild" level, 5% at the "moderate" level, and 2% at the "severe" level.

TABLE 6 WIC CLIENTS ALL 4 SITES PHQ-4 Depressive Severity Scores for WIC Sites – December 2010					
Scoring	**0-2**	**3-5**	**6-8**	**9-12**	**TOTAL**
Risk Level	**None/Minimal**	**Mild**	**Moderate**	**Severe**	
Total	1158	150	63	28	1399
Percentage	83%	11%	5%	2%	100%

For comparison purposes, at the same WIC visit, each client was also administered the PHQ-9. Staff randomized the two screeners so that some clients initially completed the PHQ-4 followed by the PHQ-9, while other clients initially completed the PHQ-9 followed by the PHQ-4. We randomized the administration of the two separate screeners in order to minimize order bias. We then compared scores with the results for the PHQ-9 by site. Because the PHQ-9 was the standard screener given to WIC clients since May 2010, over the month, some women only turned in the PHQ-9 rather than both screeners. Thus 165 clients did not have data for both surveys, but only turned in one. Table 7 shows the results for all 1564 WIC clients who returned a PHQ-9 survey:

TABLE 7 PHQ-9 Depressive Severity Scores for WIC Sites - December 2010					
SCORING	**0-4**	**5-9**	**10-14**	**15-27**	**Total**
RISK LEVEL	**None/Minimal**	**Mild**	**Moderate**	**Severe**	**N**
BRENTWOOD WIC Clients	167	33	17	7	224
Percent	75%	15%	8%	3%	224
CONCORD WIC Clients	313	62	18	12	405
Percent	77%	15%	4%	3%	405
PITTSBURG WIC Clients	452	73	32	19	576
Percent	78%	13%	6%	3%	576
RICHMOND WIC Clients	296	39	18	6	359
Percent	82%	11%	5%	2%	359
ALL WIC CLIENTS	**0-4**	**5-9**	**10-14**	**15-27**	**TOTAL**
Total	1228	207	85	44	1564

Percentage	79%	13%	5%	3%	1564

The next set of tables demonstrates the level of compatibility between the PHQ-4 and PHQ-9 among the sample WIC population.

We begin by viewing the PHQ-4 and PHQ-9 risk levels by geographic WIC site. The data is presented by percentage for each risk level with total numbers completing each of the screeners as well. Of note, these tables only give a broad picture of how well each screener performed at each site. They cannot be directly compared, since these tables include a number of WIC clients who *only* answered one screener. The comparisons for the women who completed *both screeners* will be presented after this set of tables. The results comparing the all of the women who completed either one or both of the screeners are displayed on Tables 8-11:

Table 8

BRENTWOOD	None-Minimal	Mild	Moderate	Severity	TOTAL
PHQ-4	83%	11%	3%	3%	209
PHQ-9	75%	15%	8%	3%	224

Table 9

CONCORD	None-Minimal	Mild	Moderate	Severity	TOTAL
PHQ-4	84%	9%	5%	2%	376
PHQ-9	77%	15%	4%	3%	405

Table 10

PITTSBURG	None-Minimal	Mild	Moderate	Severity	TOTAL
PHQ-4	85%	12%	5%	2%	493
PHQ-9	78%	13%	6%	3%	576

Table 11

RICHMOND	None-Minimal	Mild	Moderate	Severity	TOTAL
PHQ-4	82%	12%	3%	2%	301
PHQ-9	82%	11%	5%	2%	359

Table 12

AVERAGE	None-Minimal	Mild	Moderate	Severity	TOTAL
PHQ-4	84%	11%	5%	2%	1379

PHQ-9	79%	13%	5%	3%	1564

Across the four geographic sites, the percentages for "none-minimal" range from 75%-82% for the PHQ-9 and from 82%-85% for the PHQ-4. Table 12 above presents the average for the four sites. The PHQ-4 shows 18% of the women screening in for "mild to severe" depression or anxiety, while the PHQ-9 shows 21% of the women screening in for "mild to severe" depression (Note: the PHQ-9 screens for depression, while the PHQ-4 screens for depression or anxiety). Overall, the two screeners appear to have good comparability with approximately 3% of those screening in on the PHQ-9 being missed by the shorter screener.

Comparability of the PHQ-4 and PHQ-9

To measure how well each screener performed for the same group of women, we focused on a smaller, yet still sizable, sample of women from the total (N=1564) who completed *both of the screeners* (N=1389). We analyzed the degree to which the two screeners showed women screening at the same risk levels. We also separated out the women who did not have comparable screening results if the reason for the difference was due to the fact that they screened in on the PHQ-4 for "anxiety" (see Table 13). Since the PHQ-9 does not screen for anxiety while the PHQ-4 does, it was important to separate out this group. However, since one goal of the past year WIC project was to include "anxiety" as an important component of their mental health screener, it is useful to show the additional number of women who are identified by the PHQ-4 for anxiety, yet who are *not* identified by the PHQ-9. Table 13 displays the results by site and across all sites.

Table 13
WIC Clients by Site and Across All Sites
Comparability of the PHQ-4 and PHQ-9

Site	Compatible		Incompatible due to anxiety		Incompatible		TOTAL WIC Clients	
	N	%	N	%	N	%	N	%
Brentwood	155	82%	3	2%	31	16%	189	100%
Concord	308	82%	17	6%	50	12%	375	100%
Pittsburg	462	85%	27	5%	55	10%	544	100%
Richmond	236	84%	14	5%	31	11%	281	100%
TOTAL	**1161**	**84%**	**61**	**4%**	**167**	**12%**	**1389**	**100%**

Findings indicate that the two screeners have very good to excellent comparability. The comparable range for the PHQ-4 and PHQ-9 placing women in the same risk level categories is 84% with a range from 82% to 85%. Table 13 also indicates that 4% of women who screened in on the PHQ-4 for "anxiety" would be missed if the only screener was the PHQ-9. Overall 12% of the screeners proved to

be incompatible, with a range from 10-16%. Of note, these are fairly small numbers by site, and may vary if further comparisons were conducted over a longer period of time with larger numbers.

The main task of this latter component for this report was to compare the two screeners to see how consistent the findings were for the two populations, and to provide some guidance and recommendations about using the PHQ-4 with the WIC population.

Summary and Recommendations

Overall, both the PHQ-4 and PHQ-9 screeners worked as intended, and are useful in selecting out those at-risk women in need of referrals, brief interventions and other supportive emotional or mental health services. The PHQ-4 only contains four questions, two on anxiety and two on depression, and qualifies as an ultra-short screener. This quality is of benefit in a busy WIC environment.

In summary, given this limited analysis, the findings on the comparability of the two instruments indicate that the PHQ-4 will work as well as the PHQ-9 with this specific WIC population. It is difficult to ascertain what the specificity and sensitivity of the PHQ-4 is with the WIC perinatal population without a larger, more in-depth, well-designed study that follows women who are screened for emotional health issues and includes outcome measures.

As the literature review demonstrates, the validation of all mental health screeners for perinatal populations is still in its infancy. The major problems with the validation studies conducted to date are:

1. Lack of sufficient sample size (Of note, the WIC sample in CC County is quite large)
2. Lack of diverse racial/ethnic groups in the samples (While not included in the data collected for the WIC sample for this small analysis, the WIC population in CC County is quite diverse)
3. Absence of a diagnostic interview or screener validation by other methods or outcomes (DSM-IV, SF-20, MH professional diagnosis, etc.)
4. Good specificity on most screeners but poor to fair sensitivity (i.e., sensitivity = proportion of patients with a condition/disease [in this case, depression] who test positive or "true positives"; specificity = proportion of patients without the condition/disease who test negative or "true negatives")
5. There are only a few validation studies on the PHQ-4, although there are a fair number of studies on the PHQ-9 which forms the basis for the PHQ-4 and PHQ-2

The following recommendations emerge from the evaluation:

- The PHQ-4 shorter perinatal depression and anxiety screening tool has shown through this analysis to have very good to excellent comparability with the longer PHQ-9. The shorter

version is beneficial to both clients and staff, both of whom are pressed for time during WIC classes and appointments, and so can be recommended as a replacement for the PHQ-9.

- As stated in the earlier perinatal depression report, but worth repeating, ideally mental health screening should be undertaken throughout the prenatal and perinatal period, as well as at child wellness visits.

- Also ideally, a uniform instrument should be adopted countywide. However, it is not apparent from the literature review of validation studies on such screeners with perinatal populations that *any one* short screener stands out above the rest. For example, although much advocacy and a number of studies have been conducted on the Edinburgh depression screener, the findings still indicate that it needs, as do ALL screeners, more study with outcomes that can demonstrate good sensitivity and specificity, thus the need for follow-up diagnostic data when evaluating the screener's performance.

- Finally, continuing to use the Contra Costa Perinatal Depression to Wellness Network as a resource will prove beneficial. For the WIC perinatal depression project, it is important to stress that there is a need for outcome measures to be included in the future. Ideally, an evaluation study would be undertaken that would follow women who were screened and referred, identify those who do have as well as those who do not have a mental health diagnosis at follow-up from the screening, if a treatment plan was undertaken among those diagnosed, what services they received, and most importantly if they improved. The PDWN notes in their Vision/Mission Statement that *all* components for assessing perinatal depression are needed; the screener is merely one tool, specifically: *" …and through our efforts we will strengthen, link, and provide screening, prevention, intervention and referral services to promote wellness."*