What big size you have! Using effect sizes to determine the impact of public health nursing interventions

K. E. Johnson[1]; B.J. McMorris[2]; L.A. Raynor[3]; K. A. Monsen[2,4]

[1]The University of Texas at Austin, School of Nursing; [2]University of Minnesota, School of Nursing; [3]University of Minnesota, Division of General Pediatrics and Adolescent health, Department of Pediatrics, Medical School; [4]University of Minnesota, Institute for Health Informatics

Keywords

Nursing informatics, public health nursing, methods

Summary

Background: The Omaha System is a standardized interface terminology that is used extensively by public health nurses in community settings to document interventions and client outcomes. Researchers using Omaha System data to analyze the effectiveness of interventions have typically calculated *p*-values to determine whether significant client changes occurred between admission and discharge. However, *p*-values are highly dependent on sample size, making it difficult to distinguish statistically significant changes from clinically meaningful changes. Effect sizes can help identify practical differences but have not yet been applied to Omaha System data.

Methods: We compared *p*-values and effect sizes (Cohen's *d*) for mean differences between admission and discharge for 13 client problems documented in the electronic health records of 1,016 young low-income parents. Client problems were documented anywhere from 6 (*Health Care Supervision*) to 906 (*Caretaking/parenting*) times.

Results: On a scale from 1 to 5, the mean change needed to yield a large effect size (Cohen's $d \geq 0.80$) was approximately 0.60 (range = 0.50 – 1.03) regardless of *p*-value or sample size (i.e., the number of times a client problem was documented in the electronic health record).

Conclusions: Researchers using the Omaha System should report effect sizes to help readers determine which differences are practical and meaningful. Such disclosures will allow for increased recognition of effective interventions.

Correspondence to:

Karen E. Johnson, PhD, RN
The University of Texas at Austin, School of Nursing
1710 Red River, Austin, TX 78701
Email: kjohnson@mail.nur.utexas.edu

Appl Clin Inform 2013; 4: 434–444
DOI: 10.4338/ACI-2013-07-RA-0044
received: July 21, 2013
accepted in revised form: September 14, 2013
published: September 25, 2013
Citation: Johnson KE, McMorris BJ, Raynor LA, Monsen KA. What big size you have! Using effect sizes to determine the impact of public health nursing interventions. Appl Clin Inf 2013; 4: 434–444-
http://dx.doi.org/10.4338/ACI-07-RA-0044

1. Background

Data from electronic health records (EHRs) are useful for evaluating the impact of public health nursing interventions on client outcomes. Despite long-standing calls for researchers to calculate effect sizes as the final step in hypothesis testing [1–3], many continue to report only *p*-values. Because *p*-values are highly influenced by sample size [4], meaningful clinical changes may be masked; small sample sizes that yield statistically non-significant results may, in fact, be clinically important, whereas large sample sizes that yield highly statistically significant results may over-exaggerate the practical impact of an intervention. This is particularly relevant when evaluating data from EHRs where some interventions are more frequently used than others, yielding both large and small sample sizes (i.e., the number of times a client problem was documented in the electronic health record) that can mask the impact of interventions if only statistical significance is reported. Effect sizes, if reported, can help identify clinically meaningful changes and enable practitioners to make evidence-based decisions when developing care plans for their clients.

1.1 Omaha System

The Omaha System, a standardized interface terminology, has been used extensively in community care settings for documentation of public health nurses' (PHN) assessment, intervention, and evaluation of clients across a variety of settings (e.g., maternal-child home visits, hospice). The Omaha System has three components: the Problem Classification Scheme to document assessment findings, the Intervention Scheme to document interventions, and the Problem Rating Scale for Outcomes to document client outcomes [5]. Specifically, the Problem Rating Scale for Outcomes is used in conjunction with the Problem Classification Scheme (i.e., assessment findings classified into 42 standardized problems) to evaluate client progress [5].

1.2 Use of Omaha System for Client Outcomes Research

To evaluate the effectiveness of interventions on client outcomes, Omaha System researchers examine aggregate data documenting changes in clients' Knowledge, Behavior, and Status (KBS) across the entire program for each problem documented in the EHR. Each area of the Problem Ratings Scale for Outcomes (i.e., Knowledge, Behavior, Status) is rated on a standardized scale from 1 (e.g. No knowledge) to 5 (e.g., Superior knowledge). Ratings are reliable across problems; for example, a one-point difference for the Pregnancy problem is the same as a one-point difference for the Mental Health problem. This can result in very large sample sizes for problems that are documented frequently and very small sample sizes for problems that are not commonly experienced by clients in the program. For example, across maternal child health programs, Growth and Development is a frequently documented problem (documented 67–1011 times across agencies), Neglect is a less frequently documented problem (documented 4–269 times across agencies), and Sanitation is rarely documented [6]. On the five-point Likert-type scales used to document Knowledge, Behavior, and Status, a one point change (e.g., from 2 ["minimal knowledge"] to 3 ["basic knowledge"]) may be clinically meaningful. Yet, statistical significance of change will not be reached if the problem is rarely documented (i.e., small sample size), and the impact of the intervention may go unrecognized.

1.3 Statistical Approaches

Despite widely identified limitations to null hypothesis testing, researchers continue to exclusively report only *p*-values [2]. Reasons why researchers continue to rely exclusively on *p*-values include experience and familiarity with the procedure, and the intuitive and dichotomous interpretation of the *p*-value (i.e., $p < 0.05$ suggests meaningful results) [2]. However, one of the most important criticisms of *p*-values relates to power: studies with large sample sizes will be overpowered to detect even the smallest of differences as statistically significant [2, 4, 7]. In addition, researchers are less likely to submit studies with non-significant *p*-values for publication, thus resulting in publication bias and potential under-reporting of results that may have practical meaning for practitioners [2].

When reported alongside measures of statistical significance, effect sizes help research consumers determine the practical meaning of results [2, 4, 7, 8]. Effect sizes represent a collection of standardized and unstandardized indices that describe the magnitude of differences between means and the strength of associations among variables [4]. Cohen's *d* (independent samples t-tests), R^2 (simple and multiple regression), and Cohen's *f* (analysis of variance) are just a few examples of effect sizes researchers may use to aid in the interpretation of study results [4]. Cohen's *d* is an appropriate measure of effect size to use for Omaha System research, in which paired *t*-tests are calculated to determine differences in mean KBS scores between admission and discharge. Cohen's *d* can be calculated using the pooled standard deviation and the means of the two samples [4]. Because effect sizes are estimates derived from a sample, researchers are also encouraged to report confidence intervals with effect sizes [1, 9]. Yet among studies that have answered the call for effect size reporting, confidence intervals are infrequently reported [9].

2. Objectives

Using a sample of parents enrolled in a public health nurse home visitation program, the purpose of this study was to demonstrate differences between *p*-values and Cohen's *d* in describing clinically meaningful changes in KBS scores for Omaha System problems. In doing so, we highlight the relationship between statistical significance and sample size and provide preliminary effect size benchmarks for practice. We compared *p*-values and effect sizes (Cohen's *d*) for mean KBS differences between admission and discharge for 13 client problems to describe magnitude of changes in KBS outcomes recorded during public health nurse (PHN) home visits to high-risk parents.

3. Methods

This descriptive study employed data from an existing EHR containing KBS scores for 1,016 young low-income parents (mean age 23 years; 98% female; 20% Hispanic; 32% African American, 32% European American, 23% Asian/Pacific Islander, and 13% Other race) discharged from a Midwest PHN agency in 2009. The mean length of services received was 322 days (median 223; range 2–2954).

3.1 Sample

The analytic sample consisted of KBS scores for 13 Omaha System problems documented anywhere from 6 (Health Care Supervision) to 906 (Caretaking/Parenting) times, with a mean of 4.2 problems per client (median 4; range 1–13). Omaha System KBS outcomes are Likert-type ordinal scales which are PHN-documented observational measures of KBS (i.e., Knowledge, Behavior, Status) relative to Omaha System problems. Scores range from 1 (lowest) to 5 (highest). Reliability and validity of the instrument was established through federally funded research [10].

3.2 Variables

Thirteen problems from the Problem Classification Scheme were documented in the EHR and served as our variables for the study (▶Table 1). PHNs used the Problem Rating Scale for Outcomes to score each client problem. The Problem Rating Scale for Outcomes consists of three five-point, Likert-type ordinal scales to measure the entire range of severity for the dimensions of *Knowledge*, *Behavior*, and *Status* relative to each client problem. *Knowledge* is defined as what the client knows; *Behavior* as what the client does; and *Status* as the severity of the client's signs/symptoms [5]. Each of the five-point subscales is a continuum providing an evaluation framework for examining problem-specific outcomes (1 = lowest to 5 = highest). Definitions of the ratings for each scale are as follows: *Knowledge* (none, minimal, basic, adequate, or superior knowledge); *Behavior* (never, rarely, inconsistently, usually or consistently appropriate); *Status* (extreme, severe, moderate minimal, or no signs/symptoms) [5].

3.3 Analysis

Statistical significance and effect sizes for KBS difference scores were calculated for each Omaha System problem (▶Table 2). *P*-values were calculated using the paired samples t-test procedure in SPSS 14; the cutoff for significance was set at $p<0.05$. Effect sizes (Cohen's d) and 95% confidence intervals were calculated using a SAS effect size macro [11], using SAS v9.2 (SAS Institute, Cary, NC). Although there is no consensus on what magnitude of effect is necessary to establish clinical significance, Cohen's guidelines (1992) for interpreting effect size are as follows: small (0.2), medium (0.5), and large (0.8) [8]. These cutoffs are intended to serve as a general guideline for interpreting Cohen's *d*, rather than rigid indicators of clinical significance. Cohen's *d* is a function of the size of the mean difference, sample size, and the correlation between the paired scores.

4. Results

Outcomes were categorized into four groups based on natural breaks in sample size (6 – 24, 102 – 173, 247 – 307, and 559 – 906 documented problems) in this dataset. Mean KBS differences and effect sizes were plotted on a scatter plot (▶Figure 1). The trend line, calculated in Excel 2007, displays mean KBS changes needed to achieve large effect sizes across sample sizes. The dotted horizontal and vertical lines demonstrate intersections between the range of mean differences noted in our dataset and small, medium, and large effect sizes. On a scale from 1 to 5, the mean change needed to yield a large effect size (Cohen's $d \geq 0.80$) was approximately 0.60 (range = 0.50 – 1.03) regardless of *p*-value or sample size (i.e., the number of times a client problem was documented in the electronic health record).

Statistical significance and effect sizes of KBS change differed by sample size (▶Table 2). Four outcomes with small sample sizes and medium or large effect sizes did not achieve statistical significance: *Behavior* score for *Health Care Supervision* (n = 6; $p = 0.08$; $d = 1.07$ [95% CI: –0.07–2.22]), *Status* score for *Grief* (n = 7; $p = 0.20$; $d = 0.48$ [95% CI: -0.35 - 1.13]), *Status* score for *Health Care Supervision* (n = 6; $p = 0.36$; $d = 0.43$ [95% CI: –0.35 - 1.13]), and *Behavior* score for *Grief* (n = 7; $p = 0.36$; $d = 0.39$ [95% CI: –0.21 - 1.16]). Two statistically significant outcomes for large sample sizes had small effects: *Status* score for *Residence* (n = 270; $p< 0.0001$; $d = 0.21$ [95% CI: 0.11 - 0.32]) and *Status* score for *Caretaking/parenting* (n = 906; $p<0.0001$; $d = 0.15$ [95% CI: 0.10 - 0.21]).

5. Discussion

Findings demonstrate differences between *p*-values and Cohen's *d* in describing clinically meaningful changes in KBS scores, highlight the relationship between statistical significance and sample size, and provide preliminary effect size benchmarks for practice. Compared to Cohen's *d*, *p*-values may either exaggerate the magnitude of client change in Omaha System data or mask the impact of interventions.

5.1 Summary of findings

The mean KBS change score needed to achieve a large effect size ($d \geq 0.8$) was approximately 0.60 (range = 0.50 – 1.03), regardless of the *p*-value. Medium effect sizes ($d = 0.3 – 0.7$) were achieved with a mean score difference of 0.40 (range = 0.14 – 0.69). Small effect sizes ($d = 0.0 – 0.2$) were achieved with a mean KBS difference of 0.15 (range = 0.12 – 0.34).

All mean KBS differences for sample sizes larger than 100 were highly statistically significant, ($p \leq 0.001$). Only KBS differences of 0.50 or greater resulted in medium-to-large effect sizes. For example, the difference for *Knowledge* for the *Residence* scores was 0.79 with a corresponding $p<0.0001$ and a large effect size of $d = 1.05$. By comparison, the difference for *Status* scores for the *Residence* problem was 0.24, with a corresponding $p<0.0001$, however *d* was much smaller at 0.22. Conversely, differences for KBS problem/scales with sample sizes of less than 10 were either marginally significant or did not achieve statistical significance. Yet large effect sizes characterized *Know-*

ledge and *Behavior* scores for *Health Care Supervision* ($p = 0.04$, $d = 1.33$; $p = 0.08$, $d = 1.07$, respectively) and *Knowledge* scores for *Grief* ($p = 0.05$, $d = 1.43$).

5.2 Application to Omaha System research

For decades, experts across disciplines have encouraged researchers to report effect sizes as the final step in hypothesis testing to help determine which differences are practical and meaningful, regardless of statistical significance [2–4, 7, 8, 12]. Effect sizes are particularly relevant for describing changes in Omaha System KBS scores, due to wide ranges in problem frequencies across clients. Omaha System researchers typically have reported and compared KBS mean difference scores without a standardized way of interpreting the magnitude and importance of client change [6]. As previously mentioned, ratings of client problems are standardized across problems (e.g., a one-point difference for Mental Health is the same as a one-point difference for Mental Health). But for problems that are documented less frequently (e.g., Abuse), the same clinically meaningful changes may not achieve statistical significance.

5.3 Limitations

Effect sizes are useful for differentiating practical significance from statistical significance; yet as an estimate of a population parameter, the accuracy of an effect size estimates depends on the width of its accompanying confidence interval. Omaha System problems occurring infrequently (i.e., small sample size) have wide confidence intervals, and the true effect may be anywhere within that confidence interval. Results from the current study are therefore a starting point, not an endpoint, for establishing accurate metrics of the effect of PHN interventions. Future directions for research include replication with other sources of Omaha System data, meta-analyses of multiple Omaha System datasets, and bringing Omaha System users together to determine meaningful effect sizes for future studies.

It should be noted the metric developed with this sample of young mothers may not generalize to other groups of clients served by practitioners using the Omaha System. For example, hospice clients are not likely to experience improvements in their *Status* scores, so changes (or lack thereof) would warrant a different interpretation to determine the impact of interventions. In the future, researchers analyzing Problem Classification Scores from Omaha System data should report effect sizes for changes that occur between admission and discharge in order to establish metrics for practitioners working with other client populations.

6. Conclusions

Given increasing economic strains that threaten the funding of public health programs serving those in greatest need of services, it is more important than ever for public health practitioners to provide quality evidence supporting the need for their services. Reporting effect sizes is an important step in providing rigorous evidence to guide and financially support practice.

Clinical Relevance Statement

This study contributes an important metric that allows researchers and practitioners to look beyond *p*-values to a standardized clinically-meaningful measure. KBS differences associated with small, medium, and large effects in this sample of young parents may not generalize to other populations. In the future, we recommend that researchers analyzing KBS outcomes should report effect sizes, as well as *p*-values, in order to establish effect size benchmarks for other client populations.

Human Subjects

No human subjects were involved in the preparation of this manuscript. The Institutional Review Boards at both the University of Texas at Austin and University of Minnesota determined the study did not meet requirements for human subjects research, and therefore exempted the study from review.

Conflicts of Interest

The authors are informatics specialists and/or statisticians with expertise in use of the Omaha System in education and research. All authors declare no conflict of interest in the preparation of this manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the authors' employers.

Acknowledgments

Center for Nursing Informatics, University of Minnesota School of Nursing, Omaha System Partnership for Knowledge Discovery and Health Care Quality and the Minnesota Omaha System Users Group.

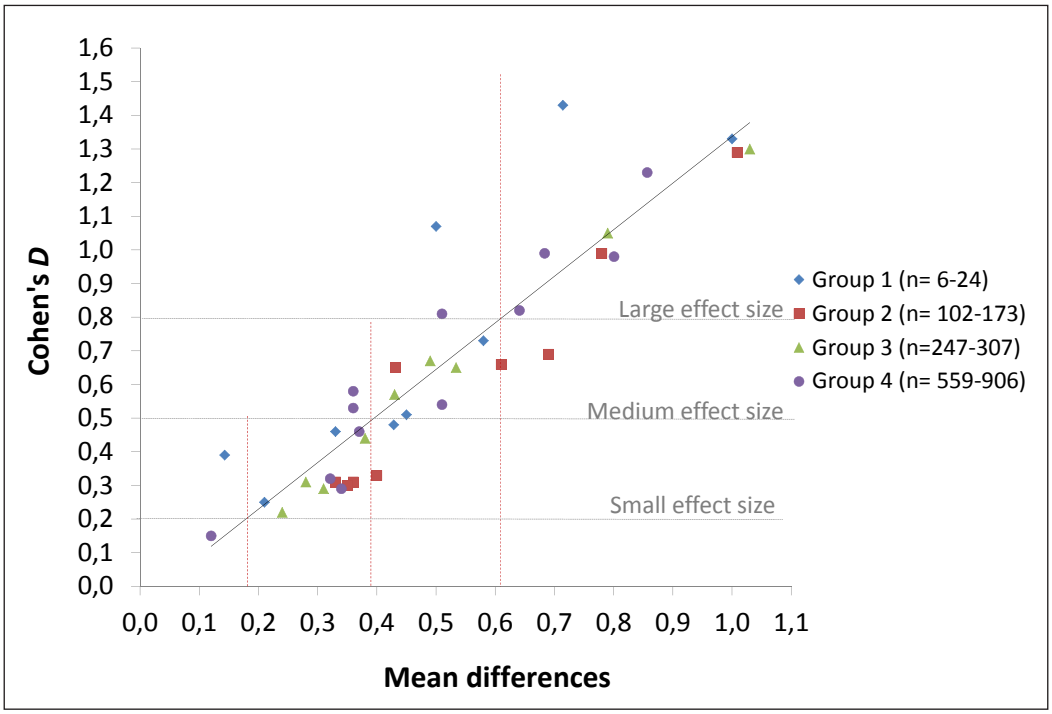

Fig. 1 Scatter plot of effect sizes (Cohen's *d*) for mean difference scores between admission and discharge

Table 1 Definition of Omaha System problems documented for study sample

Problem	Definition[1]
Abuse	Child or adult subjected to non-accidental physical, emotional, or sexual violence or injury
Caretaking/Parenting	Providing support, nurturance, stimulation, and physical care for dependent child or adult
Cognition	Ability to think and use information
Communication with Community Resources	Interaction between the individual/family/community and social service organizations, schools, and businesses in regard to services, information, and goods/supplies
Family Planning	Practices designed to plan and space pregnancy within the context of values, attitudes, and beliefs
Grief	Suffering and distress associated with loss
Health Care Supervision	Management of the health care treatment plan by health care providers
Income	Money from wages, pensions, subsidies, interest, dividends, or other sources available for living and health care supervision
Mental Health	Development and use of mental/emotional abilities to adjust to life situations, interact with others, and engage in activities
Postpartum	Six-week period following childbirth
Pregnancy	Period from conception to childbirth
Residence	Living area
Substance Use	Consumption of medicines, recreational drugs, or other materials likely to cause mood changes and/or psychological/physical dependence, illness, and disease

[1]Martin KS. The Omaha System: A key to practice, documentation, and information management. 2nd ed. St. Louis: Elsevier; 2005

Table 2 Effect Sizes (Cohen's *d*) and Statistical Significance (*p*-value) for KBS Mean Difference Scores by Problem and Scale, from Largest to Smallest Effect Size

Problem	Scale[1]	N size	Group[2]	Mean diff	p-value[3]	Cohen's d[4]	95% CI-Lower	95% CI-Upper
Grief	K	7	1	0.71	0.0465	1.43	0.27	2.60
Health care supervision	K	6	1	1.00	**0.0409**	**1.33**	0.07	2.59
Pregnancy	K	307	3	1.03	<0.0001	1.30	1.13	1.47
Communication w/ community resources	S	116	2	1.01	<0.0001	1.29	1.01	1.56
Postpartum	K	559	4	0.86	<0.0001	1.23	1.11	1.35
Health care supervision	B	6	1	0.50	**0.0756**	**1.07**	-0.07	2.22
Residence	K	270	3	0.79	<0.0001	1.05	0.87	1.22
Caretaking/parenting	K	906	4	0.68	<0.0001	0.99	0.91	1.08
Communication w/ community resources	K	116	2	0.78	<0.0001	0.99	0.75	1.24
Family planning	K	704	4	0.80	<0.0001	0.98	0.89	1.07
Income	K	706	4	0.64	<0.0001	0.82	0.73	0.91
Postpartum	B	559	4	0.51	<0.0001	0.81	0.71	0.92
Cognition	K	24	1	0.58	0.0012	0.73	0.29	1.18
Abuse	K	102	2	0.69	<0.0001	0.69	0.48	0.90
Pregnancy	B	307	3	0.49	<0.0001	0.67	0.53	0.81
Substance use	K	173	2	0.61	<0.0001	0.66	0.51	0.82
Mental health	K	247	3	0.53	<0.0001	0.65	0.51	0.89
Communication w/ community resources	B	116	2	0.43	<0.0001	0.65	0.42	0.88
Postpartum	S	559	4	0.36	<0.0001	0.58	0.49	0.68
Residence	B	270	3	0.43	<0.0001	0.57	0.42	0.72
Family planning	B	704	4	0.51	<0.0001	0.54	0.46	0.62
Caretaking/parenting	B	906	4	0.36	<0.0001	0.53	0.46	0.60
Cognition	S	24	1	0.45	0.0243	0.51	0.07	0.94
Grief	S	7	1	0.43	**0.1996**	**0.48**	-0.35	1.13
Income	B	706	4	0.37	<0.0001	0.46	0.37	0.54
Health care supervision	S	6	1	0.33	**0.3632**	**0.46**	-0.44	1.35
Pregnancy	S	307	3	0.38	<0.0001	0.44	0.32	0.55
Grief	B	7	1	0.14	**0.3559**	**0.39**	-0.21	1.16
Abuse	S	102	2	0.40	0.0008	0.33	0.13	0.54
Income	S	706	4	0.32	<0.0001	0.32	0.25	0.38
Mental health	B	247	3	0.28	<0.0001	0.31	0.19	0.44
Substance use	S	173	2	0.36	<0.0001	0.31	0.16	0.44
Abuse	B	102	2	0.33	0.0004	0.31	0.13	0.49
Substance use	B	173	2	0.35	<0.0001	0.3	0.16	0.46
Family planning	S	704	4	0.34	<0.0001	0.29	0.22	0.36

Table 2 Continued

Problem	Scale[1]	N size	Group[2]	Mean diff	p-value[3]	Cohen's d[4]	95% CI-Lower	95% CI-Upper
Mental health	S	247	3	0.31	<0.0001	0.29	0.17	0.40
Cognition	B	24	1	0.21	0.3071	0.25	-0.22	0.72
Residence	S	270	3	0.24	**<0.0001**	**0.22**	0.11	0.32
Caretaking/parenting	S	906	4	0.12	**<0.0001**	**0.15**	0.10	0.21

Notes:
[1]K = Knowledge, defined as "ability of the client to remember and interpret information; scores range from 1 (no knowledge) to 5 (superior knowledge)"
B = Behavior, defined as "observable responses, actions, or activities of the client fitting the occasion or purpose; scores range from 1 (not appropriate behavior) to 5 (consistently appropriate behavior)"
S = Status, defined as "condition of the client in relation to objective and subjective defining characteristics; scores range from 1 (extreme signs/symptoms) to 5 (no signs/symptoms)" [5]
[2]Sample size: Group 1 (6 – 24); Group 2 (n = 102 – 173); Group 3 (n = 247 – 307); Group 4 (559 – 906)
[3]Medium and large effect sizes with statistical non-significance are bolded and boxed for illustration. Small effect sizes with statistical significance are bolded and boxed for illustration
[4]Cohen's d interpretation: 0.2 = small; 0.5 = medium; 0.8 = large

References

1. American Psychological Association. Publication manual of the American Psychological Association. 6th ed. Washington, DC: The Association; 2001.
2. Greenwald AG, Gonzalez R, Harris RJ, Guthrie D. Effect sizes and *p* values: what should be reported and what should be replicated? Psychophysiology 1996; 3(2): 175-183.
3. Huck SW. Reading statistics and research. 5th ed. Boston: Pearson/Allyn & Bacon; 2008.
4. Kotrlik JW, Williams HA. The incorporation of effect size in information technology, learning, and performance research. Inf Technol Learn Perform 2003; 21(1): 1–7.
5. Martin KS. The Omaha System: a key to practice, documentation, and information management. Reprinted 2nd ed. Omaha, NE: Health Connections Press; 2005.
6. Monsen KA, Fulkerson JA, Lytton AB, Taft LL, Schwichtenberg LD, Martin KS. Comparing maternal child health problems and outcomes across public health nursing agencies. Matern Child Health J 2010; 14(3): 412-421.
7. Zakzanis KK. Statistics to tell the truth, the whole truth, and nothing but the truth: formulae, illustrative numerical example, and heuristic interpretation of effect size analyses for neuropsychological researchers. Arch Clin Neuropsychol 2001; 16: 653-667.
8. Cohen J. A power primer. Psychol Bull 1992; 112(1): 155-159.
9. Thompson B. What future quantitative social science research could look like: confidence intervals for effect sizes. Educ Res 2002; 31(3): 25–32.
10. Martin KS, Norris J, Leak GK. Psychometric analysis of the Problem Rating Scale for Outcomes. Outcomes Manag Nurs Pract 1999; 3: 20-25.
11. Kadel RP, Kip KE. A SAS macro to compute effect size (Cohen's *d*) and its confidence interval from raw survey data. Proceedings of the Annual Southeast SAS Users Group Conference; 2012, paper SD-06.
12. Vacha-Hasse T, Nilsson JE, Reetz DR, Lance TS, Thompson B. Reporting practices and APA editorial policies regarding statistical significance and effect size. Theor Psychol 2000; 10(3): 413-425.