Never too Late to Refresh!
Qualitative and Quantitative Research for Academic Nurse Educators
FACILITATORS

- Barbara J. Patterson, PhD, RN, ANEF, FAAN
- Darrell Spurlock, Jr. PhD, RN, NEA-BC, ANEF

©2020
Copyright Notice

Absolutely no part of this presentation can be recorded, copied, or distributed without the written consent of the speaker.

No photography of the speaker or the speaker’s slides can be taken without the permission of the speaker.
Four dynamic core values permeate the NLN and are reflected in our work.

- Caring
- Integrity
- Diversity
- Excellence
Objectives

➤ Increase knowledge of best practices in the generation of evidence for nursing education

➤ Apply research principles and frameworks to research in nursing education
Conceptualizing a Research in Nursing Education Study
Characteristics of Strong Research Proposals

Addressing a significant nursing education problem
What makes a problem significant?

- Prevalence (preventable medication errors occur >7.1 million times annually)
- Extent of damage caused by the problem (>250,000 deaths annually)
- Cost ($21 billion)
Relevant and Persuasive

Questions to ask?

- Is the problem clear?
- Who is target population?
- Is this a significant problem?
- Is it clear how to fix the problem?
- Is the idea novel/innovative?
- Is the solution feasible?
- Is the idea compelling to identified stakeholders?
Challenge to construct studies that are innovative and

- Fill a gap in the literature
- Have an impact
- Produce useful findings
- Can be broadly disseminated and remain effective across sites
- Value seen by everyone at each level
Pulling it all together exemplar: National FNP Clinical Education Study

- **Innovative**: Technology-enabled recruitment for large data set
- **Relevant**: Shrinking quality clinical sites and preceptors
- **Persuasive**: Current model not sustainable nor effectively preparing NPs for practice; new knowledge adds to science
- **Collaborative**: Clinical and pedagogical expertise; design, statistical and technical expertise; design, IRB, and measurement expertise

©2020
Challenges

- Nursing Education
  - Practice linkage
  - Recognition of the science of nursing education
  - Support
Challenges

- Research
  - Outcomes, outcomes, outcomes!!!!
  - Funding
  - Human subjects concerns
    - Research or standard educational procedure
    - Student or participant
    - Quality assurance/improvement or research

©2020
Where to Start?

- Start planning early
- Conceptualize your study; know what you will need
- Thorough review the literature
- Consult the experts in the education practice area - build a research team
Common Pitfalls with Conceptualization of Study

- Does not add anything new
- Review of the literature not comprehensive and/or current
- Appears to be quality improvement project for a specific course or school
Research Designs: The Foundation for Research
What is research?

The Common Rule defines research as “a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge” (45 CFD 46.102 (d))
Research vs. QI in Teaching

- Generalizable knowledge for the profession or improving own students’ learning?
- Using own students/classroom or clinical setting as subjects/research environment
- Scientific rigor/evidence-based
What is an example of research in nursing education versus quality improvement in nursing education?

Think, Pair, Share
Refreshing Our Thinking about Research
<table>
<thead>
<tr>
<th>Quantitative</th>
<th>Qualitative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Design</strong></td>
<td><strong>Data analysis</strong></td>
</tr>
<tr>
<td>➢ Pre-test / post-test</td>
<td>➢ Categories and Themes</td>
</tr>
<tr>
<td>➢ Pilot studies</td>
<td></td>
</tr>
<tr>
<td><strong>Measurement</strong></td>
<td>Qualitative synthesis</td>
</tr>
<tr>
<td><strong>Significance – Statistical vs Practical</strong></td>
<td></td>
</tr>
</tbody>
</table>
Rigorous Designs
Pretest, posttest designs

- The most common design used in studies not answering primarily correlational research questions.
- Most commonly used design to examine intervention effects
- Period between pre- and posttest ranges from minutes to years.
Pretest, posttest designs

- Primary weakness is the numerous threats to internal validity inherently problematic with the design:
  - History
  - Maturation
  - Testing
  - Instrumentation
  - Regression to the mean
Pretest, posttest designs

• What are the requirements for researchers to make an argument for causality in their research?
  • The cause (intervention) must precede the effect (outcome)
  • Differing levels of the cause produces differing levels of the effect
  • The observed effect (outcome) cannot be plausibly explained by confounding variables, even those that are unmeasured

• Why should we move away from this type of design?
  • This design does not create evidence that can be used for causal inference because it cannot exclude the effects of confounding variables.
Pretest, posttest designs

• Why is it important to move away from pretest/posttest design in nursing education?
  • Numerous empirical reviews have found that pretest/posttest designs overestimate effect sizes by anywhere from 50-400%.

• Thus, in addition to not meeting a core requirement to argue that an intervention – and the intervention alone – caused an observed effect, the effect that is observed is likely to be some multiple of the true effect.
Pilot studies: What they are and are not

- Widely used term, with widely different meanings
  - A “mini” version of a larger study?
  - A small, underpowered, single-site study?
  - A trial of something new, prior to long-term adoption?
  - A study focusing on the feasibility and acceptability of study procedures and interventions?
Pilot studies: What they are and are not

• Starting in the 1960’s, the concept of a *pilot study* emerged in response to intervention studies that for a variety of reasons, failed to produce usable, trustworthy results.
  
  • High attrition rates/low recruitment
  • Study procedures that worked well on paper but not in real-life settings
  • Miscalibration between researchers and subjects
Pilot studies: What they are and are not

• The need to *pilot test* study procedures for feasibility and acceptability, especially when experimental designs dominated and little was actually known about a particular topics, gave birth to the *pilot study*

• Feasibility: recruitment methods, recruitment rate, study procedures, length of time required, unclear instructions or confusion, etc.

• Acceptability: time vs. value for the participant, willingness to enroll, willingness to complete, etc.
Pilot studies: What they are and are not

- Over time, the focus shifted away from feasibility and acceptability to studies seeking *preliminary evidence* of intervention effectiveness.
- This requirements for rigor are the same no matter how a study is labeled.
- Thus, pilot studies that use restricted sample demographics, small sample sizes, and focus primarily on null hypotheses statistical testing (NHST), that is, *p* values, go beyond their original conceptualization.
Qualitative Data Analysis
Themes vs Categories: What is the difference?

- What is a category?
- What is a theme?
- How does one differentiate?
- Can a theme be a category?
- How does one develop a theme or category?
What is a category?

- “...is a collection of similar data sorted into the same place...this enables the category to be defined...” (Morse, 2008, p. 727).

- Categories are important for determining the what is in the data.

What is a theme?

- an extended phrase or sentence that provides not only a unit of data, but offers interpretive, insightful ways to make sense of the data. It has been suggested that theming consists of extracting verbatim significant statements from the data and formulating meanings about them through the researcher's interpretations.

- “…is a meaningful “essense” that runs through the data” (Morse, 2008, p. 727).

What is a theme?

“themes are abstract (and often fuzzy) constructs that link not only expressions found in texts but also expressions found in images, sounds, and objects” (Ryan & Bernard, 2003, p. 87).

Where do themes come from?

Examples – Category or theme?
How do you know a theme when you see one?

➢ Education
➢ Support
Exercise

“Lately I have been wondering if I expect too much from my son…..”

Read the paragraph and write one to two themes that you derive from the passage. Share with your table.

How did the theme(s) arise?

What was your process?

Measuring Educational Concepts

Darrell Spurlock, Jr. PhD, RN, NEA-BC, NEA-BC
Measurement:

Models and Theories
• **Construct**: an abstract mental representation that is often inferred / inferred to exist based on observable events, behaviors, or other phenomena (DeVellis, 2012)

• Sometimes used interchangeably with terms like *phenomena, concepts, or variables.*

• Vary in level of abstraction
• **Latent trait**: term used by researchers to describe characteristic or trait that is *unobservable* (using existing methods of measurement) but which can be *inferred* based measures of closely related constructs
  - Examples: agreeableness, readiness, introversion, resilience, and even knowledge

• Latent trait measures must demonstrate evidence of validity and reliability in order to support claims that the latent trait is being measured.

• The claims a latent trait measure makes should be clear, explicit, and situated within a theory about the relationships between concepts/constructs.
• **Measurement**: an attempt to represent a construct using an *indicator* of the construct.

• **Measures**: indicators are formalized into *measures*, often *scales* or *instruments*.

• **Items**: measures are usually comprised of *items* that when considered together with other items, form *scores*.

• **Scores**: scores are the basis upon which inferences about the latent trait are made.
• Polit and Yang (2016) describe three main types of measures, based on the measure’s primary purpose:
  • To *predict*
    • Used to make decisions based on the likelihood of future events
  • To *discriminate*
    • Used to classify subjects into groups based on characteristics or scores
  • To *evaluate*
    • To evaluate the effectiveness of interventions or treatments

• Other organizational schemas have been suggested:
  • Thorndike and Thorndike-Christ (2011): *instructional, selection, placement, classification, and personal.*
Lacking a well-accepted taxonomy, I suggest that there are four main types of measures relevant to nursing education research (Spurlock, 2016):

- **Self-report**: a method of measurement that relies on subjects to accurately report/respond to items on a measure.

- **Observation**: a method that involves independent observation by another (usually the researcher) with recording of the data on a standardized rubric, checklist, or scoring sheet.

- **Psychometric**: a method involving gathering of responses to a number of items, frequently statements, where the construct being measured is not explicitly clear to the respondent. Response are statistically combined to infer the amount of the underlying latent trait.

- **Tests**: a method where respondents must provide the correct answer (frequently from among incorrect options) to questions or other cognitive challenges.
<table>
<thead>
<tr>
<th>Self-Report</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moral Distress Questionnaire (MDQ; Eizenberg, Desivilya, &amp; Hirschfeld, 2009)</td>
<td>Hostility Coding Scale (Loucks &amp; Shaffer, 2014)</td>
</tr>
<tr>
<td>Nurse Practitioners Evidence-Based Practice Survey (Melnyk et al., 2008)</td>
<td>Clinical Judgement Rubric (Lasater, 2006)</td>
</tr>
<tr>
<td>Evidence-based Practice Questionnaire (Upton &amp; Upton, 2006)</td>
<td>Surgical Procedure Feedback Rubric (SPR; Toprack et al., 2015)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Psychometric</th>
<th>Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-Exam Self-Efficacy Measure (Cleary et al., 2015)</td>
<td>NCLEX-RN®; NCLEX-PN®</td>
</tr>
<tr>
<td>Activism Orientation Scale (AOS; Corning &amp; Myers, 2002)</td>
<td>Basic Knowledge Assessment Test (BKAT) – various versions (Toth, 2005)</td>
</tr>
</tbody>
</table>
• Measurement models help guide instrument development, and testing – which is a research endeavor at its base. The two main theories to consider: classical test theory (CTT) and item response theory (IRT).

• CTT is the most commonly used and best understood in nursing education. The equation used to summarize CTT is:

\[ X = T + e \]

where \( X \) is the true score, \( T \) is the observed score, and \( e \) represents measurement error.

• We are most familiar with CTT terms like internal consistency reliability and reliability coefficients, such as Cronbach’s alpha.
• One challenge with CTT is that the error term is a composite term, where the sources of the error are not differentiated.

• Error resulting from test construction problems, sample characteristics, changes in scores across time, etc. are only modestly separable.

• Measurement error is generally examined through the lens of reliability, which is limited often to internal consistency reliability (Cronbach’s α).

• Many measures are developed on an atheoretical basis, without serious consideration as to which measurement model would best guide development.
• Item response theory (IRT) is an approach to measurement where both the trait being measured and the characteristics of the respondents are modeled on the same scale, a logarithmic scale.

• There are multiple techniques and models categorized under IRT, including 1-, 2-, and 3+ parameter models. The Rasch model is one form of 1-parameter IRT.

• An illustrative example of how IRT works differently from CTT:
  
  > Assume we weigh a bag of apples and the scale indicates 10 pounds. And then we weigh a bag of oranges and it too indicates 10 pounds. We have the same amount of fruit present in both bags.

  In this way, the nature of the things being measured becomes less important than the scale used to do the measuring.
Validity and Reliability
• Reliable measures perform in consistent, predictable ways.

• Ideally, when the amount of a variable (or trait) changes, the measure can detect that change and reflect it in the score (i.e., the measure is sensitive).

• Reliability is, in essence, the extent to which a measure accurately detects presents a change in the amount of the underlying variable or trait being measured.

• Another way to think of it:

\[
\text{Reliability} = \frac{\text{Signal}}{\text{Signal} + \text{Noise}}
\]
Frequently used indicators of reliability:

- **Cronbach’s α**: Internal consistency of score responses; limited because not all measures *should* be internally consistent. (KR-20 is a special case for binary items)

- **Intra-class correlation coefficient (ICC)**: ANOVA-based technique to compare the variance accounted for vs. not accounted for by a measure.

- **Split-half reliability**: Index of reliability calculated by examining the internal consistency of \( \frac{1}{2} \) of the items on a measure when compared with the other \( \frac{1}{2} \) of the items.

- **Test-retest reliability**: Calculated by comparing the correlations between measurements when the trait or variable being measured should be stable over time. (The time interval matters here.)

- **Inter-rater reliability**: The ability of a measure to produce similar measurements when the measures are conducted by different raters.
Factors affecting the reliability of scales:

- Consistency among the parts of a measure (or exam)
  - Scales or exams with sections that are highly correlated one with another tend to be more reliable than those where intercorrelations are smaller.

- Length of the scale or exam
  - Longer scales / exams tend to be more reliable because as scale length increases, the ratio of true score variance to error variance decreases, and this has the result of increasing reliability estimates.
Factors affecting the reliability of scales:

- Heterogeneity of the study sample
  - Reliability tends to be higher in heterogeneous groups and more lower in homogenous groups

- Because of this, we must always remember that reliability coefficients are properties of test scores, not of the tests themselves.
  - That is, reliability estimates are highly sample and context dependent.
“Whereas reliability concerns how much a variable influences a set of items, validity concerns whether the variable is the underlying cause of item covariation.” (DeVillis, 2012, p. 59)

“In terms of validity, the set of items themselves is neither valid nor invalid. Similarly, the scores derived from the 48 items are neither valid nor invalid. However, the authors’ interpretations of the scores might be valid or invalid. Are the authors correct in interpreting scores on the set of 48 items in terms of planfulness, organization, and determination? Thus, validity is about the accuracy or legitimacy of one’s interpretations of a test’s scores.” (Furr & Bacharach, 2013, p. 198)
Three commonly accepted “types” of validity – though other definitions exist:

- **Content validity**: The extent to which a set of items measures the construct of interest.
- **Construct validity**: The extent to which the score from a measure interacts in a theoretically-consistent way with other variables. Messick (1995) suggests construct validity is the overarching form of validity.
- **Criterion validity**: The extent to which a scale predicts (or is predicted by) an empirically important outcome or variable.
Establishing validity evidence:

- **Content validity index (CVI)**
  - Agreement by experts
- **Relationships with other variables**
  - Convergent & divergent validity
  - Predicting future events with accuracy
- **Obtaining theoretically congruent / expected scores from known groups.**
  - Ex: Higher knowledge scores among senior vs. sophomore students.
• Validity and reliability are not static properties of an instrument. They must be continually evaluated over time and across different groups.

• Whereas reliability focuses mostly on a particular set of test (scale) scores, validity deals mostly with the arguments and interpretations of those using a scale or measure.

• Statistical evidence (such as factor analysis) alone is not sufficient evidence of validity.
The Nursing Student Stress Scale is not a valid and reliable...
Evaluating Instruments and Instrument Reporting
• Barry et al. (2014) evaluated 967 articles published in health education or behavior journals between 2007 and 2010.
  • Using the journal title to group the papers, between 40-93% of papers did not report validity information and 35-80% of papers did not report reliability information.
• Many instruments used in nursing education (and other areas of nursing) have limited psychometric data available…and the data are often old.
  • Often, only the primary/first report on the measure/scale is cited by authors using the tool in later studies.
  • Apart from new reports of reliability coefficients (when appropriate), many instruments never undergo additional validity testing or revision.
• The types of measures and instruments useful to the nursing education researcher varies widely, making appraisal of measure or instrument quality even more difficult.

• Most measurement and scale development texts provide sound procedures for developing, testing, and using knowledge tests and attitude questionnaires.

• Less clear guidance is available for developing observational measures or measures based on non-standard measurement theories.
• Under- or non-reporting of validity and reliability evidence for new measures.
• Conflating dispositional or attitudinal constructs with ability or action constructs. Classic case: critical thinking.
• Using proxy measures when objective measures could or should be used.
• Treating self-reports as interchangeable with more objective measures. Example: Asking subjects to rate their knowledge vs. giving them test questions.
• Translating instruments into new languages without considering how the underlying theory may be impacted by this practice.
• Creating new instruments without an appropriate underlying theory or measurement theory.
• Failing to use best practices for item writing or construction; form design, etc.
Zell and Krizan (2014) meta-analyzed data from 357,547 subjects comparing self-assessments with objective measures across a range of tasks. The overall correlation was $r = .29$ (.11).

<table>
<thead>
<tr>
<th>Table 2. Stem and Leaf Display of Meta-Analytic Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stem</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>.6</td>
</tr>
<tr>
<td>.5</td>
</tr>
<tr>
<td>.4</td>
</tr>
<tr>
<td>.3</td>
</tr>
<tr>
<td>.2</td>
</tr>
<tr>
<td>.1</td>
</tr>
<tr>
<td>.0</td>
</tr>
<tr>
<td>Table 4. Domain and Task Moderators</td>
</tr>
<tr>
<td>------------------------------------</td>
</tr>
<tr>
<td><strong>Moderator</strong></td>
</tr>
<tr>
<td>Performance domain</td>
</tr>
<tr>
<td>Language competence</td>
</tr>
<tr>
<td>Academic ability</td>
</tr>
<tr>
<td>Intellectual ability</td>
</tr>
<tr>
<td>Sports ability</td>
</tr>
<tr>
<td>Vocational skills</td>
</tr>
<tr>
<td><strong>Medical skills</strong></td>
</tr>
<tr>
<td>Memory ability</td>
</tr>
<tr>
<td>Nonverbal skills</td>
</tr>
<tr>
<td>Academic discipline</td>
</tr>
<tr>
<td>Language</td>
</tr>
<tr>
<td>Education</td>
</tr>
<tr>
<td>Sports science</td>
</tr>
<tr>
<td>Psychology</td>
</tr>
<tr>
<td>Management</td>
</tr>
<tr>
<td><strong>Medicine</strong></td>
</tr>
<tr>
<td>Task objectivity</td>
</tr>
<tr>
<td>Objective test</td>
</tr>
<tr>
<td>Subjective test</td>
</tr>
<tr>
<td>Task familiarity</td>
</tr>
<tr>
<td>High familiarity</td>
</tr>
<tr>
<td><strong>Low familiarity</strong></td>
</tr>
<tr>
<td>Task complexity</td>
</tr>
<tr>
<td>Low complexity</td>
</tr>
<tr>
<td><strong>Medium complexity</strong></td>
</tr>
<tr>
<td>High complexity</td>
</tr>
</tbody>
</table>
Mean EKAN Score by Highest Earned Nursing Degree (N = 764)

- Diploma in Nursing: 8.86
- Associate’s degree: 9.52
- Bachelor’s degree: 10.69
- Master’s degree: 11.00
- DNP: 11.00
- PhD/DNS: 18.00

\( r = .237, p < .001 \)

Mean EKAN Score by EBP Care Confidence Level (N = 764)

- Strongly Disagree: 8.90
- Disagree: 7.44
- Neither Agree nor Disagree: 9.44
- Agree: 9.72
- Strongly Agree: 9.81

\( r = .06, p = .059 \)

Measurement of Self-Efficacy

As noted in the introduction, tests of a theory require valid assessment of its key constructs. There are serious problems in the measurement of self-efficacy in the studies under discussion. Before commenting on the validity of the measures in question, I will review briefly the standard procedure for constructing psychometrically sound scales of perceived self-efficacy. Self-efficacy is concerned with people’s beliefs in their capabilities to produce given attainments (Bandura, 2006b). All too often, this belief system is treated as though it is a generalized trait. In fact, people differ in their efficacy, not only across different domains of functioning but even across various facets within an activity domain. Consequently, there is no single all-purpose measure of self-efficacy with a single validity coefficient.

Issues identified:

• Using the wrong response scale, such as a Likert-type agreement scale where a discontinuous bipolar response is possible.
• Using language in scale items that departs from the underlying theory, such as: I will..., confidence, resilience.
• Using a restricted response scale: Does only 3-4 response options provide sufficient granularity in defining the strength of one’s beliefs?
• Changing the meaning of what’s being measured by departing from the source theory.

General Self Efficacy Scale

<table>
<thead>
<tr>
<th>Question/Item</th>
<th>Factor 1 (Proficiency)</th>
<th>Factor 2 (Altruism)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I can always manage to solve difficult problems if I try hard enough.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. If someone opposes me, I can find the means and ways to get what I want.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. It is easy for me to stick to my aims and accomplish my goals.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. I am confident that I could deal efficiently with unexpected events.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Thanks to my resourcefulness, I know how to handle unforeseen situations.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. I can solve most problems if I invest the necessary effort.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. I can remain calm when facing difficulties because I can rely on my coping abilities.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. When I am confronted with a problem, I can usually find several solutions.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. If I am in trouble, I can usually think of a solution.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. I can usually handle whatever comes my way.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As of today, how confident are you that you can:

<table>
<thead>
<tr>
<th>Question/Item</th>
<th>Factor 1 (Proficiency)</th>
<th>Factor 2 (Altruism)</th>
</tr>
</thead>
<tbody>
<tr>
<td>22. Manage therapeutic interventions safely (e.g., drainage tubes)?</td>
<td>.722</td>
<td>.104</td>
</tr>
<tr>
<td>23. Prepare clients for diagnostic procedures and treatments (e.g., colonoscopy)?</td>
<td>.699</td>
<td>.136</td>
</tr>
<tr>
<td>12. Complete your assessments in a timely manner</td>
<td>.591</td>
<td>-.091</td>
</tr>
</tbody>
</table>


"It should be noted that the construct of self-efficacy differs from the colloquial term "confidence." Confidence is a nondescript term that refers to strength of belief but does not necessarily specify what the certainty is about. I can be supremely confident that I will fail at an endeavor. Perceived self-efficacy refers to belief in one's agentive capabilities, that one can produce given levels of attainment. A self-efficacy assessment, therefore, includes both an affirmation of a capability level and the strength of that belief. Confidence is a catchword rather than a construct embedded in a theoretical system. Advances in a field are best achieved by constructs that fully reflect the phenomena of interest and are rooted in a theory that specifies their determinants, mediating processes, and multiple effects. Theory-based constructs pay dividends in understanding and operational guidance. The terms used to characterize personal agency, therefore, represent more than merely lexical preferences." See *Self-Efficacy: The Exercise of Control*, 1997, p. 382
BANDURA’S INSTRUMENT
TEACHER SELF-EFFICACY SCALE

This questionnaire is designed to help us gain a better understanding of the kinds of things that create difficulties for teachers in their school activities. Please indicate your opinions about each of the statements below by circling the appropriate number. Your answers will be kept strictly confidential and will not be identified by name.

Efficacy to Influence Decision making

How much can you influence the decisions that are made in the school?

1  2  3  4  5  6  7  8  9
Nothing Very Little Some Influence Quite a Bit A Great Deal

How much can you express your views freely on important school matters?

1  2  3  4  5  6  7  8  9
Nothing Very Little Some Influence Quite a Bit A Great Deal

Efficacy to Influence School Resources

How much can you do to get the instructional materials and equipment you need?

1  2  3  4  5  6  7  8  9
Nothing Very Little Some Influence Quite a Bit A Great Deal

Instructional Self-Efficacy

How much can you do to influence the class sizes in your school?

1  2  3  4  5  6  7  8  9
Nothing Very Little Some Influence Quite a Bit A Great Deal
Arthritis Self-Efficacy Scale

For each of the following questions, please circle the number that corresponds to how certain you are that you can do the following tasks regularly at the present time.

1. How certain are you that you can decrease your pain quite a bit?

   Very uncertain  1  2  3  4  5  6  7  8  9  10 Very certain

Items (using the same format as above):

1. How certain are you that you can decrease your pain quite a bit?
2. How certain are you that you can keep your arthritis or fibromyalgia pain from interfering with your sleep?
3. How certain are you that you can keep your arthritis or fibromyalgia pain from interfering with the things you want to do?
4. How certain are you that you can regulate your activity so as to be active without aggravating your arthritis or fibromyalgia?
5. How certain are you that you can keep the fatigue caused by your arthritis or fibromyalgia from interfering with the things you want to do?
6. How certain are you that you can do something to help yourself feel better if you are feeling blue?
7. As compared with other people with arthritis or fibromyalgia like yours, how certain are you that you can manage pain during your daily activities?
• Instrument development is often more complex and more time consuming than other forms of research.
  • Extensive literature reviews are often required.
  • Enlisting the help of experts to review items is time-intensive.
  • Sample size requirements are often higher, so recruitment can take longer and incentives are often needed.
  • The analytic methods needed are often outside the comfort zone of even experienced researchers.
• To adequately report on the essential aspects of instrument development, a longer manuscript may be required…and publishers are pushing for shorter – not longer – manuscripts.
Data Collection, Management, and Analysis

Darrell Spurlock, Jr. PhD, RN, NEA-BC, NEA-BC
Key Questions and Considerations

• Where will your data come from?
• How will you transform data from one form to another, if needed?
• Have you selected the right statistical tests and other analytic methods?
• Do you have the correct software to do the analysis?
• Do you need the services of a data analysis consultant / statistician? Do you have access to one?
Sources of Data in Nursing Education Research

• Data are collected in-person by the researcher or a member of the researcher’s team
• Data are collected in a distributed manner, by others who are assisting the PI
• Data are collected directly by the PI and/or research team using electronic data collection tools (e.g., online surveys)
• The researcher has access to or plans to use existing data (e.g., large datasets, institutional data, etc.)
Data Collected In-person

• Traditional method used to collect questionnaire, interview, or observational data.

• Team members must be trained in order to ensure data integrity and study internal validity.

• Suggestion: Do not propose a multi-site interventional study without a strong coordinating/training component for remote site study team members.

• Scheduling and logistics can be complicated and must be considered ahead of time. (Suggestion: Get letters of support for grant applications; don’t assume study sites will help you due to universal altruism)

• You must ensure you have the proper permissions, administrative and ethical, before you can start data collection. Speak to this in your proposals, and even better, get some of this work done prior to grant submission.
Data Collected In-person

- Be prepared and have all forms, supplies, and required equipment on hand.
- Be respectful of the time of study subjects and others (e.g., remote site faculty) when on-site.
- Do not depart from your approved procedures and methods (e.g., do not add other questionnaires, change the method of administration, etc.)
- Ensure any required technological resources needed for data collection are in place, such as working wi-fi, open computer labs, etc.
- If you aren’t collecting the data yourself, be available by mobile phone in case there are questions or issues. Your research team members may not know how to proceed or respond when issues arise.
Data Collected in a Distributed Manner (by remote collaborators)

- This idea sounds good, and can work, but it takes about 5x more planning and energy than most PIs allocate for upfront.
- When remote site research team members get busy, your study will be down-prioritized.
- IRB issues can also be thorny, and your remote collaborators may be unfamiliar with how to move this forward at their sites.
- I strongly recommend against distributed multisite intervention study designs (where the intervention will be delivered by someone other than the primary PI or the PI’s team) for novice researchers. This has to do with issues of intervention fidelity and the internal validity of study findings. Most nursing education studies are not funded to sufficient levels for robust multi-site distributed intervention designs.
Collecting Data Directly, in Multi-site or Distributed Designs

• Online survey and electronic data collection technology improvements have enabled this approach to be more affordable, accessible, and effective than in the past.

• Data can be collected using online survey tools (e.g., Qualtrics®, SurveyMonkey®, etc.), or via platforms like SMS (text messaging) and video (via smart phone or video conferencing).

• The primary research team remains responsible for the bulk of data collection and logistics; some assistance may be provided by remote collaborators.
Collecting Data Directly, in Multi-site or Distributed Designs

• Ideal approach if the team has the appropriate technical skills to deal with the data collection process and workflow.

• Technology can make some things easier but some things more complicated and time consuming. Plan ahead and expect to invest more time up front on design and testing than you might initially expect to need.

• Be aware of data security requirements and legal/ethical requirements in re: collecting data from human subjects in online platforms. Not just any free survey tool will do. The tool should support features like encryption, strong protection against data loss, strong (>99.99%) uptime promises, and meet all industry standards current at the time of data collection.

• Individualized study invitations via email or SMS aren’t possible if you don’t have a list of potential subjects to invite, so consider how you will obtain – or if you CAN obtain such a list.
Collecting Data Directly, in Multi-site or Distributed Designs

• Consider the work that remote sites do to facilitate your data collection efforts, and if that work deserves some sort of incentive or stipend payment to the remote site (which can help with engagement).

• Evaluate marketing best practices when considering how you will interact with potential subjects via electronic means only.

• **THERE IS EVIDENCE ABOUT WHAT WORKS BEST TO RECRUIT AND ENGAGE REMOTE PARTICIPANTS IN RESEARCH.** Learn this literature and cite it in your proposals to build trust among grant reviewers:
  • Recruitment methods; frequency of email contacts
  • Use of recruitment/study incentive payments
  • Design of survey invitations: Language, appearance.
  • Appropriate expectations of subject time/effort:incentive payment ratios.
Case Study

• Online survey of students enrolled in BSN-PhD programs throughout the US in January 2017; subjects would complete three questionnaires about Evidence-based Practice. Total time investment = 20 minutes. Recruitment goal of \( N = 700 \).

• Able to use DIRECT recruitment of subjects via personalized emails after schools provided FERPA-compliant name and email contact lists to study PIs.

• $10 Amazon.com gift code incentive delivered immediately after survey completion by email, using an automated online system.

• Recruitment: Sent 4 waves of email invitations to a population of \( N = 7,200 \) students enrolled in accredited nursing programs over the course of 3 days, Tuesday-Friday.

• Target recruitment goal of \( N = 700 \) subjects reached after just 4 days!
Collecting Data Directly, in Multi-site or Distributed Designs – Pros:

- Data collection can be efficient, quick, and the data can be re-coded/prepared by the survey/EDC platform (if desired)
- Many systems have built-in reporting/analytics capabilities so the team can see summary stats and trends as the data come in.
- Promises of data confidentiality and integrity are stronger when remote site collaborators truly don’t have access to subject data (such as when students are being studied and may be leery of providing responses due to fears about retaliation, etc.).
- Using EDC methods allows for use of data validation and enforcement: Subjects can’t input random text into an email contact box; the entered text must match the format of an email address; numerical fields can be enforced to accept no string text, etc.
Collecting Data Directly, in Multi-site or Distributed Designs: Cons

• In non-proctored contexts, more difficult to ensure that subjects meet inclusion criteria and aren’t “hacking” in some way.

• Can require sophisticated technical skills to deploy the most advanced features (like randomization of subjects to surveys, delivery of study incentives, validation checks using data from other online services, etc.)

• Survey fatigue is a major problem for many individuals and institutions. The key is in building relationships, making contact ahead of time, and reducing respondent burden as much as possible.
http://nationalnpstudy.org: A study funded by the National Council of State Boards of Nursing (NCSBN)
Using Existing Data

• Large datasets are increasingly available.
  • No hard definition, but datasets with more than 5,000 rows or 250 columns are generally considered large.

• Large datasets can frequently be handled in standard statistical software (e.g., SPSS, Stata, etc.) but you may experience issues with loading large datasets into Excel, depending on the version of Excel you use and the quality of the computer hardware you have access to.
  • Analysis or computation for large datasets is often very processor- and memory-intensive, so you must have access to adequate computer resources to do this work.

• Institutional student information systems (SIS) are one major source of “big data” that nursing education researchers may be interested in. These systems often house many academic (e.g., GPA, course grades, add/drop history, standardized test score, etc.) types of data that can be queried – often from a historical perspective.

• FERPA may apply in many instances, so check about your institution's policies on this before beginning any work with student data.
Using Existing Data

• An emerging area of interest, especially within institutions of higher education, is the use of analytics, which are mainly visual methods of data analysis, with large datasets available to institutions.

• This phenomena has implications for researchers as well, as more and more data become available, these data become a possible source for nursing education research.

• Exploratory data analysis (EDA) procedures are a type of quantitative analytic procedure being introduced in many PhD in nursing programs because of its relevance to other “big data” scenarios such as with –omics and large health system datasets.

  • Modern data learning management systems (LMS) also can produce line-by-line records of student activity, clicks, scores, and other indicators based on student engagement with the LMS. These data are generally too big to subject to periodic, standalone analysis, but are perfect for a visual analytics EDA approach.
Considerations in Transforming Data

• Frequently, the data you must analyze needs to be transformed prior to analysis.
• This can be done in a number of ways, using a variety of software and non-software tools.
• Consider that not all data derived from scales needs transformation; some of it needs only to be “entered” and then subject to analysis.
• In some cases, only basic calculations such as summing or computing a mean must be performed.
Population

Sample

Sampling Techniques

Analysis

About

Inferences
Selecting Statistical Tests

General principals:

• It is more important to be **RIGHT** with your statistical analysis plan to **APPEAR** sophisticated.
  • While there is often more than one way to conduct an analysis (in which case there may be several correct options), there are also very wrong options, too. Avoid wrong or overly complicated options.

• The statistical tests you use (or propose to use) should fully consider the following considerations (at least):
  • Assumptions of the statistical test: All inferential tests have **assumptions** about the data which can be subject to it; violating those assumptions can have more or less of an effect on the results, so be aware of the assumptions.
  • Level/type of data needed: If a technique requires a continuously measured dependent variable, as is the case with multiple linear regression, you cannot use a binary or categorical DV.
The Power of Power Analysis

- *Power* is the ability to detect a given *effect size*, given the existing sample size (*post hoc*), or to help you determine the sample size you need upfront (*a priori*) to detect the effect. Components needed to calculate sample size for a means comparison:
  - $\alpha$: probability of a Type I error (erroneously rejecting the null hypothesis)
  - Power: $1-\beta$; probability of detecting the given effect size
  - Effect size: magnitude of effect to be detected (can be based on existing knowledge or be theory-based)

- Current opinion is that power analysis is most appropriate and most useful if done *a priori*; it was developed for this purpose.
- Power analysis, done *a priori*, forces researchers to think ahead to what would be an important finding (effect size).
• The **FEWER** statistical tests you can do to answer your research questions, the better. Parsimony is highly valued in statistical analysis.
  
  • The issue here is something called **familywise error**: If you do 15 different $t$-tests, which are all NHSTs, each test will produce a $p$ value because each test has its own null hypothesis…which means that for each $t$-test, there is a certain probability, usually 5%, of committing a Type I error (rejecting the null hypothesis [saying there is a difference] when it is in fact NOT true [there is no difference]).
  
  • This means that across all your tests, you have a $15 \times .05 = .75 = 75\%$ chance of committing a Type I error among your many conclusions. There are corrections (Bonferroni’s correction is the simplest, which is to basically divide the total $\alpha$ of .05 by the number of tests you will conduct)
  
  • Even better is to only conduct the minimum number of tests needed to answer your research questions.
• Make sure the data you are collecting are sufficient in number to do the test specify in advance.
  • Many statistical tests have “assumptions” that include some aspect of sample size. You cannot do a factor analysis on a instrument with responses from only 4 subjects, for example.
  • Other frequent assumptions are normal distribution and independence of observations. Note: most data in small samples aren’t normally distributed, and most observations (i.e. data points) aren’t truly independent of one another.
• Nonparametric tests are generally used when violations of normality and/or independenc are at work.
• Make sure the data you are collecting are of the right “measurement level” to use in the analyses you plan. Age can be collected either as a number (age, in years), or as a category (18-28, 29-39, etc.). Depending on which level you use, this will change the kind of statistical analysis you can do.
  • For example, if you collect age as a continuous variable, you can use it along with another continuously measured variable, like IQ, in a Pearson’s r correlation analysis. If you collect it only in categorical form, you lose the granularity and detail in the data and are restricted to less robust techniques for statistical analysis.
  • Remember that you can always categorize data you’ve collected in continuous form; you can’t go the other way.
• Specify your main analyses *a priori*, so it does not appear that you are on a “fishing expedition” for a statistically significant *p* value with your data, especially if your research hypotheses were not supported.

• Be sure to report all the information that readers (and researchers in the future) need to fully interpret your results:
  • Don’t forget to always report SDs with your means, the test statistics with your *p* values, and, and sample sizes (if they differed for an individual test due to missing data).
  • These elements are essential for meta-analyses and replications of your work in the future.
Recognize that *p values alone generally tell us nothing of value* beyond that you had a large sample size.

- Almost every NHST will be statistically significant if the sample size is large enough.
- This is why some measure of EFFECT SIZE is supremely important to report. Examples of effect sizes include:
  - Correlation coefficients
  - Cohen’s $d$ (which indicate the number of SDs of an effect)
  - Odds Ratios
  - Relative Risk
Which statement do you find more meaningful?

1. Men who drank more than 3 alcoholic drinks per day were statistically significantly more likely to develop early cognitive dysfunction than men who drank only one drink, on average, per day ($R^2 = .17, df = 4, 198, p < .001$).

2. Men who drank more than 3 alcoholic drinks per day developed cognitive dysfunction at four times the 2.1% rate of the average man in the population ($RR = 4.1, p < .001$).
Having the Right Statistical Tools

• **SPSS®** is the most common statistical analysis program in nursing and the social sciences.
  - It can be very expensive when purchased separately, but often, universities and colleges have subscriptions to it and the cost to individuals can be $0
  - Substantial online documentation can help inexperienced data analysts navigate through new territory

• **Microsoft Excel®,** while lacking the capability for sophisticated statistical analysis, is very useful for data transformation, re-coding, and sorting functions.
  - Plugins, like XL-STAT® are available for a reasonable cost, but these tools lack the documentation and sophistication that formal stats programs like SPSS, Stata, or R have.

• **Other tools to consider:**
  - **STATA®:** Statistical analysis program that is just as powerful a SPSS, with equivalent capability and help documentation, but with more reliance on command line functionality (when compared to SPSS); output requires more “translation” into APA 6th edition than SPSS
  - **R:** Free software package that has many GUI options; requires development of at least low-level programming knowledge so that data can be manipulated at the command line.
Pre-analysis Checks

- Data formatting: Are all data correctly labeled, coded, and categorized in your analysis program?
- Data quality: examine for outliers, miscoding, duplicate cases, or other spurious values.
- Normality: use visual tools like histograms, scatter plots, and box plots.
- Missing data: How to address? Imputation, leave it missing, or omit the case?
Reporting Your Results

• Use narrative text if that is the most efficient way to convey results. (Tables/figures aren’t always better)

• Report all useful information, such as means, standard deviations, effect sizes, and if possible, confidence intervals. Report actual test statistic values, degrees of freedom, and actual p values.
  • Note: Statistical programs truncate very small p-values into “.000” – so these should be reported as <.001, not “= .000”

• When possible, report results in “meaningful units” rather than in purely statistical terms.
  • Take inspiration from the clinical example of number needed to treat (NNT) where the interpretation of the value is in response to the question: How many patients (students) would I need to treat before 1 would benefit? (Smaller values are better.)
Next steps
Where we need to Advance the Science
What do you see as the next steps in research for advancing the science of nursing education?

Think, Pair, Share
Qualitative Research Synthesis
Secondary Data Analysis

Secondary analysis is the use of an existing data set to find answers to a research question that differs from the question asked in the original or primary study (Hinds et al. 1997, p. 408).

Secondary Data Analysis: Benefits

- Result in useful findings that may not have been explicit before in the primary study
- Promote generalizability
- Re-examine data in light of different insights
- Develop theory
- Cost effective, reduce participant burden
- Facilitate more research with vulnerable populations

Secondary Data Analysis: Challenges

- Qualitative findings are never free from researcher perspective – researcher-participant relationship
- Context based on time data were collected & interpreted understanding may have changed
- Missing data from original data set
- Fit between data and methods

Secondary Data Analysis: Approaches


Syntheses: Meta-synthesis or Meta summary

- Process: begin with delineated problem and purpose
  - Retrieve all relevant qualitative studies based on inclusion/exclusion criteria
  - Set parameters on topical area, population, time period. Methodology is more problematic.
  - The follow up appraisal of selected articles.
  - Data analysis – synthesis of findings

- Meta-synthesis: creates a new integrative interpretation of findings
- Metasummary: a count of the frequency of the findings

Meta-synthesis Example


- Purpose of this article is to synthesize the current qualitative literature on challenges faced in nursing education for students with English as an additional language.

- Methodology: Synthesis of 10 qualitative studies (out of 50 results, 1996-2009)
  - Looked at key themes and issues in nursing education
  - Compared content, relationships, and common themes across all 10 studies
  - 2 categories were derived – challenges faced by EAL & reinforcements that help EAL students succeed
Evidence-Based Teaching Practice
Dissemination

Nursing education needs a robust evidence base to advance the science!

- Aspect often underappreciated by many applicants
- Creative venues, disseminate in new ways
- Show how findings may be used by stakeholders
- Innovative idea → Research brief → full manuscript
Translation of Evidence into Teaching Practice

- Barriers – “This is the way we have always taught it…”
  - Colleagues
  - Demanding workload
  - Insufficient time
  - Lack of formal educational preparation or background
  - Institutional barriers: lack of supportive leaders and mentors

- Fostering a culture of educational research
Translation of Evidence into Teaching Practice

- Integrating research findings into teaching practice
  - Systematic reviews
  - Integrative reviews
  - Meta-analysis/meta-synthesis

- Evidence for educational policy
  - NCSBN’s National Simulation Study (2014)
Reflections

Building the evidentiary base for the science of nursing education will require:

- Innovation
- Robust designs
- Funding
- Research teams with different expertise
- Large, diverse, multi-site involvement
- Researcher commitment to longitudinal and/or long-term projects
- Investment by Schools of Nursing
Recommendations

- Increase the cadre of teacher-scholars who provide leadership through the SoTL
- Creation of evidence-based teaching practice guidelines or best practices.
  - Integrative reviews and meta-analyses/synthesis
  - Develop measurement tools
- Nurse educators are conducting research, it must be available to all educators so replication and refinement can occur.
Group Activities
Identifying Challenges & Key Issues in Generating and Translating Evidence
Liberating Structures

What is made possible?

This activity is designed to help the group move from taking inventory of the facts and lessons from the last 2 days (What Happened?) to making sense of these facts (So What?), and finally to identifying the actions and next steps that flow logically from what has been learned (Now What?).

- Arrange into small groups of 5-7 people, taking paper on which to write.
- For the first 1-2 minutes, independently write out ideas about “What Happened?” – what are a few things you learned?
- Next 7-8 minutes, discuss in your group your What Happened? thoughts. Someone in the group should capture themes.
- Now, repeat the steps for So What? and Now What? (1-2 minutes individually and then 7-8 minutes as a group, capturing themes)
- Last 10-12 minutes, report out and share group findings.