1. SESSION: Organizing survey delivery

TBC

1. SESSION: Designing better surveys
2. **[8] Roadmap**
   1. [3] From latent concept of interest to measured quantity
      1. Example: Psychological scales
         1. Well validated
         2. But: Measures context dependent; transfer HIC to LIC/MIC challenging (Laajaj and Macours, 2020)
      2. Example: Economic concepts
         1. Questions often factual, and measured by one obvious item
         2. But: Even for simple measures, assuming implicitly measure is robust across time and space (and in the face of shocks)
         3. But: More complex economic concepts even closer to psychological scales
      3. Conclusion
         * Given budget constraint, we optimize on survey length
         * Need for methods to quantify measure quality (within a specific context)
         * Need for methods to design good measures to capture latent concept (within a specific context)
   2. [2] First part: Quantifying survey item quality
      1. Item quality
         1. Measure latent concept, and no other concept(validity)
         2. Measure latent concept consistently (reliability)
            1. Across time
            2. Across space
            3. (In the face of shocks)
         3. Economize survey time
      2. Classical test theory
         1. True score theory
         2. Validity
         3. Reliability
   3. [2] Second part: Accounting for realistic response behavior
      1. Attention
      2. Comprehension
      3. Retrieval
      4. Inference
      5. Editing
   4. [1] Concluding thoughts: Designing better surveys
3. **[29/32] Quantifying survey item quality**
   1. **[2/5] Classical test theory**

[1] Classical test theory provides a quantitative framework to testing the reliability and validity of a scale based on its items.

It is based on true score theory that, under certain assumptions, allows to quantify reliability

[1] Reliability

Consistency of a test score across many measures

Reliability is a property of the scores of a measure (not that of the measure itself), and thus sample dependent

[1] Ways to improve reliability

Item wording (clarity of expression)

Adding additional items to measure the same concept

Making items more similar

Item analysis

Computing item difficulties

Computing item discrimination

Rank items, and replace bad items with new items (i.e., too hard, too easy, low discrimination score)

[1] Validity

See slides

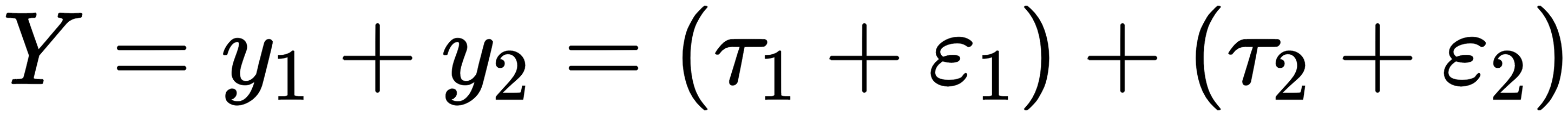
[1] Ways to improve validity

See slides

* 1. **[12] True score theory**

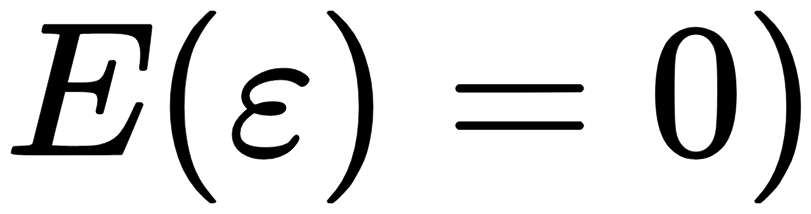
(Cronbach, 1951; Guttman, 1945; Lord & Novick, 1968; Novick & Lewis, 1967)

* + 1. [3] A two-item model
       1. Measurement model
          1. General setup

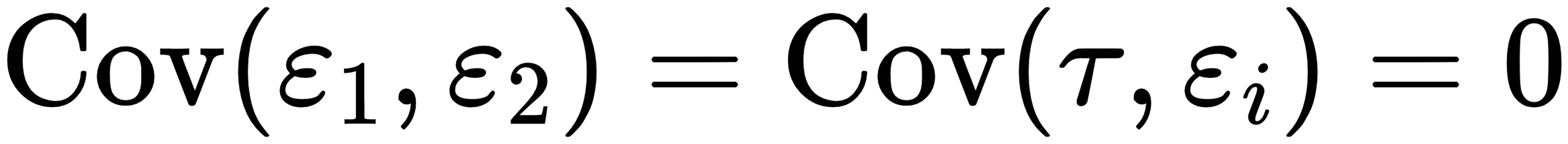
$Y=y\_{1}+y\_{2}=\left(\tau\_{1}+\varepsilon\_{1}\right)+\left(\tau\_{2}+\varepsilon\_{2}\right)$

* + - * 1. Assumption: Unbiased measure

$E(\varepsilon)=0)$

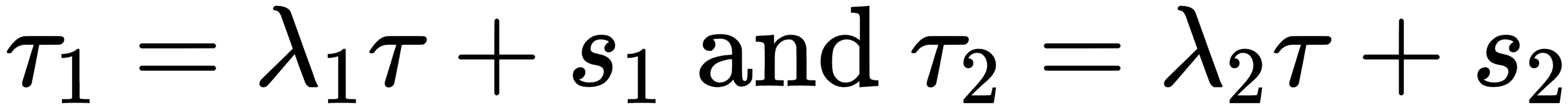


* + - * 1. Assumption: Independent errors

$\operatorname{Cov}\left(\varepsilon\_{1}, \varepsilon\_{2}\right)=\operatorname{Cov}\left(\tau, \varepsilon\_{i}\right)=0$

* + - 1. Tau transformation
         1. True score

$\tau\_{1}=\lambda\_{1} \tau+s\_{1}$ and $\tau\_{2}=\lambda\_{2} \tau+s\_{2}$



* + - * 1. Assumption: Identical latent construct

$\tau\_{1}=\tau\_{2}$



* + 1. [3] Common models in the literature
       1. Parallel measures
          1. $\tau\_{1}=\tau\_{2}$



* + - * 1. $Var(\varepsilon\_{1})=Var(\varepsilon\_{2})$
      1. $\tau$ equivalence
         1. $\tau\_{1}=\tau\_{2}$



* + - * 1. Relaxes $Var(\varepsilon\_{1})=Var(\varepsilon\_{2})$ to allow for $Var(\varepsilon\_{1}) \neq Var(\varepsilon\_{2})$
        2. Essential $\tau$ equivalence

Relaxes $\tau\_{1}=\tau\_{2}$ to allow for $s\_{1] \neq s\_{2}$ in $\tau\_{1}=\tau+s\_{1}$ and $\tau\_{2}=\tau+s\_{2}$

True scores can differ, but true-score variance is constant ($\neq$ error variance!)

* + - 1. Congeneric measures
         1. Relaxes $\tau\_{1}=\tau\_{2}$ to allow for $\lambda\_{1] \neq lambda\_{2}$ in $\tau\_{1}=\lambda\_{1} \tau+s\_{1}$ and $\tau\_{2}=\lambda\_{2} \tau+s\_{2}$
         2. As per $\tau$ equivalence: Relaxes $Var(\varepsilon\_{1})=Var(\varepsilon\_{2})$ to allow for $Var(\varepsilon\_{1}) \neq Var(\varepsilon\_{2})$
         3. True scores, true-score variance, and error variance can all differ
    1. [5] Violations of key assumptions
       1. Multidimensionality
       2. Congeneric forms

We conclude that in most applications items on scales should measure a predominant general factor as well as narrower group factors. In other words, scale items should be multidimensional, and thus violate the essential tau equivalency assumption. In practice, many scale developers claim unidimensionality for their scales. We suspect that they have ignored minor group factors in their analyses either intentionally—understanding the group factors were minor relative to the general factor—or unintentionally due to the difficulties associated with assessing the number of factors underlying item data (Hattie,1985; Zwick & Velicer, 1986).Scales that assess very narrow constructs might be unidimensional. Even for these scales, it is unlikely that the essential tau equivalency assumption would hold in that it requires the amount of item variance associated true score variability to be the same across all items on a scale. It is more likely that a congeneric model, $X\_{ij} = \gamma\_{j} F\_{i} + E\_{ij}$ would hold that permits factor loadings to vary among items.

* + - 1. Correlated errors

Green & Hershberger (2000) suggested that correlated errors occur if measurement errors on prior items affect responses to later items either directly or indirectly (through prior items).

Respondents to non-cognitive measures can choose, at some level of consciousness, to respond consistently across items and, in so doing, may allow their responses to earlier items to affect their responses to later items. They hypothesized that consistent responding is likely to be maximized if (a) items are clustered based on their wording, and (b) respondents are not sufficiently interested in the measurement process to attend to the specific wording of each item, perhaps due to the length of the total measure. Although no studies have empirically investigated the applicability of these autoregressive/moving averages models, the literature does support the effect of item order on factor analyses of non-cognitive measures (Schurr & Henriksen, 1983). In contrast, item arrangement tends to have only a minimal effect on factor analytic results of cognitive measures (Leary & Dorans, 1985).

As described in most psychometric textbooks, coefficient alpha yields spuriously high estimates of reliability for speeded tests with right–wrong answers. Rozeboom (1966) suggested that correlated errors are introduced with speeded tests.

Correlated errors might also be introduced if subgroups of items are associated with different stimulus materials (Steinberg, 2001; Steinberg & Thissen, 1996; Wainer & Kiely, 1987; Yen, 1993). Examples include comprehension items associated with reading passages, direction items associated with maps, items concerning different responses to interpersonal situations described in vignettes, and negative mood self-report items associated with different aspects of one’s life(e.g., home, work, and so on).

* + - 1. *[1] Loose ends*: Advanced approaches related to true score theory
         1. Factor models

True score model has been redefined as a factor analytic model (e.g., Bollen, 1989; Jöreskog, 1971; McDonald, 1999; Miller, 1995; Raykov, 2001; Zinbarg, Revelle, Yovel, & Li, 2005)

* + - * 1. Multidimensionality (hierarchical models)

Scales should be unidimensional for interpretability (Hattie, 1985; McDonald, 1981)

But: Likely to include additional factors (threat to essential tau-equivalence)

Some psychometricians argue structure underlying items should be more consistent with hierarchical model (or general form bifactor model; e.g., Gerbing & Anderson,1988; Humphreys,1985; McDonald,1999; Reise et al.,2000; Roznowski, Tucker, & Humphreys,1991).

Bifactor model (Chen, West, & Sousa, 2006; Rindskopf & Rose, 1988; Schmid & Leiman, 1957; Yung, Thissen, & McLeod, 1999)

Includes not only general factor $F$

But includes also $k$ group factor(s), $F\_{k}$, such that $X\_{i j}=\lambda\_{j} F\_{i j}+\sum\_{k=1}^{K} \lambda\_{j k} F\_{i j k}+E\_{i j}$

* 1. **[10] Test reliability**
     1. [2] Stability (Test-retest reliability)

[1] Consistency of test scores across repeated administrations at time t, t+1

Procedure:

Administer test to group of individuals G at time t

Re-administer test to group G at time t+1

Compute correlations between first, second set of scores (also see, item-total correlation)

Both single-item scales and multi-item scales

[1] Tau equivalence assumption

Individual’s type time-invariant

No learning

No carry-over

No reactivity

* + 1. [2-3] Parallel forms reliability

[1] Consistency of test scores across alternate test forms T1, T2 (by assumption parallel measures)

Procedure:

Develop alternate test forms equivalent in content, measurement properties, etc.

Administer test form T1 to group of individuals G and administer test form T2 to group of individuals G

Compute correlations between first, second set of scores

Alternate test forms exist for many psychometric tests, such as intelligence tests, personality scales, etc.

[1] Tau equivalence assumption

Alternate forms are parallel

Individual’s type time-invariant

No learning

No carry-over

No reactivity

Measures to limit potential learning, carry-over effects

Cross-randomizing order of T1, T2

Administering alternate test forms to two randomized groups G1 and G2

[1] Advantages:

Alleviates concerns of learning, carry-over effects

Alleviates concerns over reactivity effects

Disadvantages

Often difficult to create alternate test forms (also measurement properties should ideally be known!)

Typically impossible to guarantee alternate forms are parallel measures

* + 1. [2] Split-half reliability

[1] Consistency of test scores across two halves T\_part1 and T\_part2 of one test, treated as alternate test forms (parallel measures)

Procedure:

Split test in half

Administer one half T\_part1 to group of individuals G,

Administer other half T\_part2 to group of individuals G

Compute correlations between first, second set of scores

[1] Independence assumption

Item difficulty

Test fatigue

-> Split by odd/even items, not first/second half

* + 1. [2] Internal consistency

[1] Consistency of scores across single items within a test

Most common measure: Cronbach’s $\alpha$ (Cronbach, 1951)

Mean of coefficients of all possible split-half combinations

Generalization of Kuder-Richardson formula for binary items

$\alpha=\mathrm{K} /(\mathrm{K}-1)\left[1-\left(\sum \sigma\_{k}^{2} / \sigma\_{\text {total }}^{2}\right)\right]$

where $\mathrm{K}$ is the number of items, $\sigma\_{k}^{2}$ is the sum of the $k$ item score variances, and $\sigma\_{\text {total }}^{2}$ is the variance of scores on the total measurement (measured as per Crocker and Algina, 1986, p.95).

[1] Alpha is (roughly) the square of the correlation between true score variance and total score variance

Items perfectly uncorrelated, $\alpha = 0$

Items perfectly correlated (same sign), $\alpha = 1$

Problem of negative \alpha

$\sigma\_{k}^{2} > \sigma\_{\text {total }}^{2}$

Items perfectly correlated and mixed signs (measuring different concepts)

Negative alpha calls into question the integrity of the scores (Thompson, 2003)

* + 1. [1-2] Inter-rater reliability

See slides

Link to Item response theory

* 1. **[5] Test validity**
     1. Construct validity
        1. See slides
     2. Content validity
        1. See slides
     3. Criterion validity (concurrent, predictive)
        1. See slides

Link to Item response theory

1. **[~37] Accounting for realistic respondent behavior**

**How will this fit?**

**Scenario 1: Mapping test theory assumptions onto known biases**

**Scenario 2: Laying out the whole cognitive theory**

**Scenario 3: Clever cross-over**

* 1. **Sources of systematic measurement error 1**

General answering tendencies

Social desirability and experimenter demand effects

Question design and context

Effects of memory and attention

Question order effects

Sensitive questions

Language and translation

* 1. **[~4] Attention and survey motivation**

**General principles**

Cognitive load

Level of abstraction

Level of generality (and order effects)

Recognition versus recall

Recall period

Survey length

Contextualizing questions (catering to people’s education, background, etc)

* 1. **[~5] Question comprehension**
     1. [1-2] Lexical comprehension
        1. Concept definition
           1. Smoking a cigarette (Suessbrink, Schober, and Conrad, 2000)
        2. Context
           1. Marital vs. life satisfaction (Schwarz, Strack, and Mai, 1991)
           2. Fitness vs. illness (Couper, Tourangereau, and Conrad, 2007)
     2. [2] Semantic comprehension
        1. Translation and back translation (Dalton et al., forthcoming)
        2. Conversational vs. standardized interviews (Schober and Conrad, 1997)
     3. [1-2] Pragmatic comprehension
        1. Grice’s 5 cooperative principles
        2. Absence of success vs. presence of failure
     4. Reference periods
     5. Rating scales
     6. Open and closed answer formats
     7. Question context
     8. Translation, language, and culture
  2. **[~14] Retrieval from memory**
     1. [3] LTM capacity
        1. STM/WM vs. LTM
           1. STM/WM
           2. Types of LTM

Episodic

Semantic

* + - * 1. Recall periods
    1. Episodic and semantic LTM memory
       1. [1] Recall and recognition, distinctiveness and similarity
       2. [1-2] Organization principles
          1. Temporal and conceptual organization
          2. Encoding specificity (Godden and Baddeley, 1975),
          3. Context reinstatement (Fisher and Quigley, 1992)
       3. [1] Retrieval cues
       4. [1] Count-and-recall and estimation
       5. [1] Reconstructive view of memory
          1. “Implanted memories” and 90/10%
          2. Repetition as priming
    2. Facilitating episodic recall
       1. Reference periods
       2. Recall cues
       3. Task decomposition
       4. Temporal direction
    3. Memory accuracy
       1. [1] Confidence
          1. Confidence vs. accuracy
          2. “Feeling of knowing”
       2. [1] Accuracy markers
          1. Reaction time, etc.
          2. Role of age
  1. **[~6] Estimation and response mapping**
     1. Frequency and impression-based estimation
        1. Partial episodic recall or semantic frequency estimates
        2. Over/underestimation (Conrad)
        3. Age
        4. Availability heuristic
        5. Subjective theories
        6. Frequency scales
     2. Mapping of memory trace onto answer
        1. Response order effects
           1. Primacy
           2. Recency
           3. WM capacity
        2. Rounding of frequency estimates
        3. Context effects (Cowan, Murphy, and Wiener, 1978)
        4. Rating effects
  2. **[8] Editing for social expectations**
     1. Privacy concerns
     2. Social desirability and self-presentation
        1. Evidence for social desirability bias
           1. Face to face vs. self-admin
           2. Response time
        2. Rating scales
        3. List experiments
        4. Vignette measures
        5. Implicit Association Test

## Audio Computer Assisted Self-Interviewing

* + 1. Moral and cultural appropriateness
    2. Experimenter demand effects

1. **[3] Concluding thoughts: Designing better surveys**
   1. **General principles for good survey items**
   2. **Specific lessons**
      1. Use multi-item measures
      2. Play to respondents’ cognitive strengths (and avoid weaknesses)
      3. Mind respondents’ response biases
      4. Quantify item quality to select best items
2. SESSION: Piloting surveys and interventions (incl. qualitative methods)

See *PhDcourse\_pilot*

**Check:**

[**https://medium.com/pew-research-center-decoded/how-much-do-the-numbers-used-in-survey-scales-really-matter-227d84ab2a13**](https://medium.com/pew-research-center-decoded/how-much-do-the-numbers-used-in-survey-scales-really-matter-227d84ab2a13) **(Rating scales 0-6and 1-7 doesn’t matter much, except mid point chosen more often if number of levels even)**

[**https://mbrg.bsg.ox.ac.uk/method/surveying-sensitive-topics-implicit-association-test**](https://mbrg.bsg.ox.ac.uk/method/surveying-sensitive-topics-implicit-association-test) **(IAT)**

[**https://mbrg.bsg.ox.ac.uk/method/surveying-sensitive-topics-using-audio-computer-assisted-self-interviewing**](https://mbrg.bsg.ox.ac.uk/method/surveying-sensitive-topics-using-audio-computer-assisted-self-interviewing) **(Self-interviewing)**

**Open-ended Questions Open-ended survey questions allow respondents to answer in their own words. Open-ended questions also allow the researcher to explore ideas that would not otherwise be aired and are useful where additional insights are sought (Salant & Dillman, 1994, p. 81). They are also useful where the researcher is less familiar with the subject area and cannot offer specific response options. Open-ended questions require greater thought and contemplation on the part of the respondent, and are, therefore, more time intensive to answer (Salant & Dillman, 1994, p. 79). The results obtained from open-ended questions are also more difficult to analyze. Finally, it is more difficult to identify a single course of action from the broad range ofresponses that are received to open-ended questions (p. 80). 2.2.2.2 Closed-ended Questions In contrast, closed-ended questions require the respondent to choose from among a given set of responses (McIntyre, 1999, p. 75). Closed-ended questions with ordered choices require the respondent to examine each possible response independent of the other choices. The choices form a continuum of 2-7 responses, such as those provided by Likert scales and numerical ranges. These types of questions are easiest for respondents to answer and for researchers to analyze the data. The second type of closed-ended question is the closed-ended question with unordered choices (Salant & Dillman, 1994, p. 83). These questions ask the respondent to compare possible responses and select one. Multiple choice questions are an example of this type. The researcher must ensure that the respondent is given a comprehensive selection ofresponses. Closed-ended questions with unordered choices are useful for ranking items in order of preference. The third type of closed-ended question is the partial closed-ended question in which the respondent is asked to compare possible responses and select one, or write in “other”. Salant and Dillman (1994) observed that most respondents choose one of the given responses when this type of question is presented (p. 84). Closed-ended questions may also be categorized as: (a) questions that describe and evaluate people, places, and events; (b) questions that measure responses to ideas, analyses, and proposals; and (c) questions that measure knowledg**

**Cognitive Tasks Required for Survey Response Schwarz (1999) considered the cognitive tasks that respondents perform when asked to answer a survey question. 2-9 The first cognitive task is question interpretation. Specifically, the respondent must understand what the researcher is asking and determine what information will best meet that request (Schwarz, 1999, p. 66). The second cognitive task is response formulation. Schwarz (1999) noted that respondents tend to construct new judgments as that is less cognitively demanding than determining whether previously-held judgments meet the specific constraints of the question (p. 66). In the third cognitive task, the respondent communicates the response to the researcher. Schwarz (1999) observed that given response options may constrain cognitive activity so that the respondent only generates a response that directly fits the given options (p. 66). Additionally, the respondent may intentionally or unintentionally edit the response to meet unstated expectations of political correctness or social norms. 2.2.5 Sources of Measurement Error Salant and Dillman (1994) cautioned interviewers to avoid leading respondents to specific answers, interpreting questions for them, or projecting an image that suggests certain answers are desired (p. 19). Each is a source of measurement error. The respondent is another source of measurement error. Salant and Dillman (1994) observed that respondents may answer as they think the interviewer wants them to answer (p. 20). Additionally, responses to surveys may not reflect the true beliefs, attitudes, or behaviors of the respondents. Respondents may intentionally provide false responses to invalidate the survey’s results or choose not to reveal their true insights for a host of personal reasons, reasons that may not be rational or even understood by the respondent (Browne & Keeley, 1998, p. 114). Isaac and Michael (1997) identified three additional sources of bias associated with the respondent. First, the conduct of a survey is generally outside the daily routine of most respondents and their participation may invoke feelings of being special (p. 137). The Hawthorne effect, named after the Hawthorne Works of the Western Electric Company, is perhaps the most well-known example of this type of bias. The Hawthorne studies of worker performance in 1927 found that worker performance improved simply from the awareness that experimental attempts were being made to bring about improvement. The second type of respondent bias noted by Isaac and Michael (1997) was the propensity of respondents to agree with bias inherent in the wording of the question, such that respondents more readily agreed with positively-worded questions. Finally, respondents may give consistently high or low ratings, reflecting a rater bias that detracts from the validity of the results (Isaac & Michael, 1997, p. 137)**

**Aron, A., & Aron, E. N. (1997). Statistics for the behavioral and social sciences: A brief course. Upper Saddle River, NJ: Prentice Hall. Attewell, P., & Rule, J. B. (1991). Survey and other methodologies applied to IT impact research: Experiences from a comparative study of business computing. Paper presented at The Information Systems Research Challenge: Survey Research Methods. Bell, S. (1996). Learning with information systems: Learning cycles in information systems development. New York: Routledge. Browne, M. N., & Keeley, S. M. (1998). Asking the right questions: A guide to critical thinking. (5th Ed.). Upper Saddle River, NJ: Prentice Hall. Creswell, J. W. (1994). Research design: Qualitative and quantitative approaches. Thousand Oaks, CA: Sage. Fowler, J., Floyd J. (1995). Improving survey questions: Design and evaluation. (Vol. 38). Thousand Oaks, CA: Sage Publications. Isaac, S., & Michael, W. B. (1997). Handbook in research and evaluation: A collection of principles, methods, and strategies useful in the planning, design, and evaluation of studies in education and the behavioral sciences. (3rd Ed.). San Diego: Educational and Industrial Testing Services. Kraemer, K. L. (1991). Introduction. Paper presented at The Information Systems Research Challenge: Survey Research Methods. Levy, P. S., & Lemeshow, S. (1999). Sampling of populations: Methods and applications. (3rd ed.). New York: John Wiley and Sons. Lucas, J., Henry C. (1991). Methodological issues in information systems survey research.Paper presented at The Information Systems Research Challenge: Survey Research Methods. McIntyre, L. J. (1999). The practical skeptic: Core concepts in sociology. Mountain View, CA: Mayfield Publishing**