CS-BIGS 2(1): 9-20
http://www.bentley.edu/csbigs/vol2-1/goodman.pdf

Critical-Thinking Assessment: A Case Applying Resampling to Analyze the Sensitivity of a Hypothesis Test to Confounding

William Goodman
University of Ontario Institute of Technology, Canada

The Case for this paper illustrates how a statistical resampling technique can be used to provide a sensitivity analysis for the possible vulnerability of a hypothesis test to confounding. In spite of other methods for guarding against confounding influences, there will always be uncertainties associated with the model at the base of a hypothesis test. It is demonstrated that, by modeling those influences that might cause confounding, simulation can be used to determine the sensitivity of a hypothesis test to give a false positive, at various strengths of the confounding. To illustrate these methods, a case is introduced based on the longstanding efforts of nursing educators to improve the teaching of critical thinking (CT) in their programs, and to apply assessment tools to test whether their work has been successful. Given years of mixed results, some suggest that the measurement tools employed may be flawed. This raises the question of how sensitive the hypothesis tests they employed would have been to confounding effects, if these had been introduced by using unreliable measures to assess critical thinking.

Keywords*: Critical thinking, Resampling, Confounding, Hypothesis Tests*

Introduction to the Case

The Case for this paper illustrates how a statistical resampling technique can be used to provide a sensitivity analysis, for the possible vulnerability of a hypothesis test to confounding. A confounder is a variable, or interaction of variables, whose (unmodeled) association with the dependent variable could bias our estimate of an effect being studied (Pearl 1998). Although major causes for confounding (such as errors or changes in measurement) can be elaborated, and hopefully controlled for (Habicht *et* al 1984), there will always be assumptions and uncertainties associated with the model at the base of a hypothesis test. Sensitivity analysis (SA) helps us to attribute the uncertainty of a model's output (such as unexpected values of a test statistic) to different sources for potential uncertainty in the model's input (Saltelli 2002).

The specific case, below, arises from a longstanding call for nursing educators to add critical thinking (CT) as a core nursing competence in their curricula. According to responsible and certifying organizations such as the National League of Nursing (U.S., 2004) and the College

of Nurses of Ontario (Canada, 1999), clinical nurses, like any professionals, require strong skills in critical thinking and decision making that could impact how they meet new and unexpected situations. Responding to this mandate, nursing educators have spent years in (a) developing new ways to convey these crucial skills, and (b) developing and applying assessment instruments to test whether CT has actually improved. Many studies have been reported on their efforts, dating back to the 1970's, and employing a wide variety of contexts, interventions, and assessment tools in their protocols.

Regrettably, the collective results from all these efforts have been inconclusive, and present no clearly replicated evidence of improvement in Critical Thinking. That was the conclusion of a metastudy published by Adams (1999) covering years 1977-1995; and in Table 1, in the next section, we find similar mixed results for subsequent studies.

On thus failing to measure the expected or hoped-for critical thinking improvements with their instruments, educational institutions have often sought for internal causes for the problem. But possibly there is a more global explanation—involving the sensitivity to confounding influences of some of their commonly used testing procedures. This effect could possibly explain an aggregate pattern of random "false positives" in some studies, interspersed with negative or neutral results for others.

Confounding, as just noted, can invalidate the results of a study. Proper test design, such as using controlled experiments and designing to ensure that interaction effects will be distinguishable from main effects, can eliminate some of the causes of confounding (Box, Hunter, and Hunter 2005). But strict controls are often difficult to design, for example in an education setting where a program is either implemented (for a whole class) or it is not. Sometimes, a test may rely heavily on certain assumptions (e.g. that a commercially sold test instrument is reliable) that are not properly questioned until after the fact. In such cases, a sensitivity analysis could demonstrate how vulnerable the test results would be (or might have been) to confounding, if it exists.

But how can we measure a confounding effect whose existence and strength, if any, are not known in advance? This challenge is not impossible, as shown by the analogy of a power curve that is constructed for a hypothesis test: we cannot know the real risk of a Type II error for a test, because we do not know where the real population mean is located. Yet we can draw a chart to show the variable likelihood of such error, as the true mean moves closer or further from the null mean. In like manner, this paper demonstrates how a resampling-based simulation could be used to assess how varying degrees of confounding could impact the outcome of a test. As the confounding effect becomes stronger, we may find that test statistics which appeared at the edges of the test's null distribution, and so appear "significant", may turn out be in the non-critical regions of the true population distribution. The construction of such a simulation is described in this paper.

Background

Definition and Attributes of Critical Thinking

Valid assessment must "directly measure that which it is intended to measure" (Brown, Race, & Smith 1996). This seemingly straightforward criterion can be problematic if there is no settled definition for what is to be measured. Critical Thinking (CT) is in this category.

A definition by Robert Ennis provides the minimum elements for defining CT (1996): "Critical thinking is a process, the goal of which is to make reasonable decisions about what to believe and what to do." Yet the marks of "reasonableness" remain up for debate. Does critical thinking refer to just a skill, or to dispositions towards using the skill? Is the object of CT something that can be taught in isolation, or is it intertwined with particular areas of expertise? Some classical viewpoints are presented in Norris (1992). A nursing education perspective is offered by Edwards (2003). In nursing education, critical thinking is valued less for improving logical argumentation than for engaging reflective practice and "theories in use," as espoused by Dewey (1959), Argyris (1982), and Schon (1987).

Among efforts to build consensus on the definition and attributes of critical thinking, to provide a basis for assessment, a 1990 Delphi study directed by Peter Facione has been very influential. A diverse panel of experts in critical thinking reached consensus on statements of "core CT skills and sub-skills" as well as the "affective dispositions of critical thinking" (Facione 1990). These lists form the basis for Facione's commercially available assessment instruments. Other efforts, geared to nursing education, have been spearheaded by Gordon (2000), the National League for Nursing (2003), and Scheffer and Rubenfeld (2000).

Challenges of Critical Thinking Assessment

Over time, a variety of instruments for measuring CT have been developed and applied. In an "integrative review" of progress in the teaching of critical thinking to

students in accredited nursing programs, Adams (1999) reviewed 20 assessment studies reported from 1977 to 1995. Most, but not all, used the popular Watson-Glaser Critical Thinking Appraisal instrument, designed to assess a composite of critical thinking attitudes (e.g. of inquiry and acceptance of evidence), knowledge (e.g. of the nature of abstractions and valid inference forms), and skills (in appropriately applying the knowledge and attitudes). To her surprise, Adams reached the disturbing conclusion that taken in aggregate, the studies provide "no consistent evidence that nursing education contributes to increasing the critical thinking of nursing students".

Adams' conclusion could be easily re-stated, following many similar studies undertaken in subsequent years, many of which are summarized in Table 1, below.

Since 1995, two newer instruments have become prominent—both developed by Peter Facione and reflecting the consensus on CT reached by the 1990 Delphi Study, mentioned earlier: The California Critical Thinking Skills Test (CCTST) and the California Critical Thinking Disposition Inventory (CCTDI) (Facione, Facione, & Giancarlo (FFG) 1997; FFG 2000; FFG 2001). Yet even papers coauthored by Dr. Facione have reached mixed conclusions regarding the significance of CT differences, as detected by his own instruments.

So how is it possible that, given years of documented concern by nursing educators to impart skills and dispositions in critical thinking to their students, they can report such little consistent success? Allowing for variation in study quality, instrumentation, variables, and protocols, should there not be *some* consistent pattern of meaningful advances? If not, the problem may not lie with the teachers, courses, or students, but with a fundamental weakness in the assessment procedures being used. Among those who suggest this, Tucker (1996) argues that critical thinking is very context dependent, and measures whose virtue is to be easy to grade and universally applicable are suspect. Wilson's (2000) statistical objections to conventional assessments are especially pertinent for the present case. For applying test-retest experiments with commercially available tests, he concludes, "Test-retest reliability coefficients are relatively low, making it difficult to distinguish actual changes from background noise. The relationship between scores produced by these instruments and learner's [actual critical thinking] skills is unknown and warrants further research." This low test reliability could be one explanation for the apparently random outcomes of successive CT-measurement studies.

Specific Research Context for this Case

Wilson's suspicions about CT tests' reliability appeared to be confirmed in a 2003 study among students in a Canadian university (Goodman 2004). Incoming students wrote the CCTDI instrument within a larger survey, as part of a longitudinal study in a four-year, baccalaureate program in nursing. Fortuitously, a test-administration error was discovered in the final screens of the web-delivered survey (partly containing the CCTDI); so the researchers asked students to rewrite the last component of the survey at a later date. This effectively enabled a test-retest comparison between students' answers for the same questions, on sections that were *not* affected by the administration error, and thus identical in both versions.

Surprisingly, between the students' two response opportunities, there was little or no correlation between their individual answers to the exact same questions. This lack of consistency, *at the individual-question level*, might be called the "non-responsiveness" of the instrument. Admittedly, the time delay of several months between re-writes was not ideal for the test-retest comparison. Yet the CCTDI, in particular, claims to measure a *disposition*—which is something that should *persist*, in the absence of a systematic dispositional change. Neither a systematic change nor persistence was observed.

Note that the "non-responsiveness" of the instrument is different from an instrument's being "unreliable", as measured, for instance, by Cronbach's alpha. Alpha reliability requires questions to be grouped consistently with respect to constructed variables, collectively maintaining their correlational structure from test to re-test. An instrument could be unreliable according to alpha if question responses in a proposed cluster are not properly aligned, yet still the instrument could be *responsive—if* respondents at least answer the same *individual* questions fairly consistently upon retest. The CT-measuring instrument apparently lacked this property.

A measurement tool that is non-responsive in this sense would certainly yield inconsistent, and possibly random, results if applied in studies. But if a tool's results are virtually random, how could its output *ever* be found to have a significant correlation to some variable? This might happen rarely—as a Type I error for the hypothesis test, but it should not occur as often as have the occasional "significant" results reported from CT assessments, based on suspect instruments. The following research objectives have evolved from that question.

Table 1. Summary of Studies on Critical-Thinking Assessment

Author (Yr)	Primary Question(s)	Sample / Design	Tool	Result	+/-
Adams, Stover, & Whitlow (1999)	Compare the scores of students at two levels of their program.	Longitudinal study: 203 BSN students. Assessed in sophomore year, second semester; and senior year, final semester.	WGCTA (Watson Glaser)	No reported increase in the CT abilities.	**Neg**
Angel, Duffey, & Belyea (2000)	Compare learning outcomes in the acquisition of knowledge and the development of CT skills in relation to strategy of clinical teaching.	Quasi-experimental pretest-posttest design: 142 junior nursing students. (WGCTA given upon admission to control for variation among groups.)	Custom, case-based questionaire, with open-ended questions (two relating to CT).	CT abilities increased significantly, but with no relation to instructional approach. Other cross-sectional variables were examined, but none were predictive of changes in skills.	**Mixed**
Beckie, Lowry, & Barnett (2001)	Evaluate the attainment of CT skills of students before and after a curriculum revision of a baccalaureate nursing program.	Longitudinal Study, with three cohorts of students: n=55 before a curriculum revision; then the first two cohorts following the revision (n=55 & n=73). Each cohort was evaluated at program entry, midpoint and exit.	CCTST (California Critical Thinking Skills Test)	Cohort 2's scores were significantly higher than Cohort 1's, but the CT scores decreased for Cohort 3. (Note that Cohort 2 had entered the program with higher CT scores.)	**Mixed**
Bowles (2000)	Evaluate the relationship of CT to clinical-judgment abilities in nursing students at the completion of their program.	Compare CT scores of 65 nursing students from two baccalaureate nursing programs with their assessed clinical-judgement skills..	CCTST (California Critical Thinking Skills Test) and a clinical judgement test	Apparently significant, but weak (r2 = .04) correlation between the scales. Not all subscales correlate. Correlation found between CT and GPA.	**Very Weak Pos.**
Brown, Alverson, & Pepa (2001)	Compare the changes in CT abilities of students pursuing various pathways in the same nursing curriculum,	Comparative: n=123, comprised of traditional 4-yr stud's (n=45); RN-BSN (n=35); Accelerated (n=43)	WGCTA (Watson Glaser)	Significant CT improvements were reported for the traditional and RN-BSN groups, but not for those in accelerated programs, who enter with another degree.	**Mixed**
Colucciello (1997)	(a) Compare differences in CT skills and dispositions for different academic levels; and (b) Examine whether there's a relationship between CT skills and CT dispositions.of nursing students.	Comparative: 328 students, stratified into one sophomore group (n=94), two junior groups (n=65; n= 64); & two senior groups (n= 59; n=46).	CCTST (California Critical Thinking Skills Test) and CCTDI (California Critical Thinking Disposition Inventory)	(a) CT scores increase significantly by academic level. (b) A significant, but weak, correlation found between CT skills vs dispositions.	**(a) Pos (b) Pos**
Facione & Facione (1997)	Examine the relation between critical thinking skills and dispositions and other established measures of quality in nursing practice.	Aggregate study, based on data provided from 50 collaborating nursing programs.	CCTST and CCTDI	Mixed results, especially for dispositions.	**Mixed**
Giancarlo & Facione (2001)	Did students' critical thinking dispositions change over four years of their education; and were there significant differences between academic or demographic groups.	Longitudinal and Cross-Sectional: Overall n=1117 students in liberal arts university participated in 1992 (mostly freshmen) or 1996 (mostly seniors); n=147 participated at both stages.	CCTDI	Results based on raw scores are mixed.	**Mixed**
Magnussen, Ishida, & Itano (2000)	(a) Following a curriculum change, examines if CT skills will increase. (b) Examine the relative progress of subgroups, based on high, medium, or low initial scores.	Primarily longitudinal : During first week of school (n=228); then during final semester of program (n=257); 150 paired-scores.	WGCTA	(a) Overall, no significant increase reported between students' entry scores and final scores. (b) By strata: Initially-low scoring students increased their scores; initially high-scoring students decreased their scores. [Could this be regression to the mean?]	**(a) Neg (b) Mixed**

Martin (2002)	In the context of a clinical simulation, evaluate the relationships among CT, decision making, and clinical nursing expertise.	Stratified sample of graduates plus RN's, selected from different schols and health care agencies. (n=149).	Locally developed, case-based instrument	Correlations found between CT and professional expertise (links with Benner's models of expertise development). No correlations with CT found by nursing stream or based on demographics	**Pos.**
May, Edell, Butell, Doughty, & Langford (1999)	Compares the relationship between critical thinking skills and clinical competence	143 graduating senior nurses.	CCTST, CCTDI, and a custom clinical competence eval'n tool.	No significant, overall correlations found between CT and clinical competence. A few weak positive relationships were found between CCTDI subscales and clinical competence.	**Mixed**
McCarthy, Schuster, Zehr, & McDougal (1999)	Compares critical thinking abilities for beginning and graduating nursing students.	Cross-Sectional: sophomores (n=156) in two groups; seniors (n=85) in two groups.	CCTST and CCTDI	Finds significant increases for seniors in CT scores, compared to sophomores. Also finds a significant correlation between the two CT measures employed.	**Pos.**
Pepa, Brown, & Alverson (1997)	Evaluates the influence of an accelerated nursing curriculum on students' abilities to think critically.	Longitudinal: Two tracks: Trad'l BSN (n=45); and Accelerated (n=43). Both groups tested at beginning and end of program.	WGCTA (Watson Glaser)	The traditional students, but not the accelerated students showed increased CT scores at the end of their programs.	**Mixed**
Smith-Blair, & Neighbors (2000)	Explores and examines the possible use of the CCTDI for evaluating the disposition to CT of students entering a critical care orientation program.	Descriptive: 65 nurses from 5 hospitals, tested during 1st week of orientation	CCTDI	{There's no real test here…but the Authors' confidence in the assessment's reliability--to the extent of proposing to use results for decision making--deserves notice.}	***n/a***
Spelic, Parsons, Hercinger, et al (2001)	Following a curr'm revision, "to evaluate the dev't of CT skills in students in the BSN curr'm."	Longitudinal: 136 students in 3 prgms tested on entry and on exit: Traditional(n=51); Accelerated(n=68); and "LEAP" (1-yr program for licensed RN's) (n=17)	CCTST	Significant increases in mean scores for all groups The authors raise some score-reliability issues.	**Pos.**
Tawari (2004)	Compare the effect of problem-based learning and lecturing on the development of nursing students' critical thinking skills.	Pretest/posttest: Treatment group used problem-based learning techniques (n=40); Control used traditional lecture (n=38). . Random distribution of the students into the groups.	CCTDI (CT disp'ns) + qualitative methods	No significant findings for the quantitative study.	**Neg.**
Walsh & Hardy (1999)	Identify differences in critical thinking dispositions among college students by academic majors and by gender.	Comparative: Overall n=334, subdivided by gender, race, and major (grouped by practice vs non-practice disciplines)	CCTDI	Although some differences among groups appeared to be significant , there was no clear pattern of difference. (The picture is "somewhat muddied".)	**Mixed**

Research Objectives

(1) To demonstrate that if sample statistics are collected using a flawed measurement tool, resulting hypothesis tests may be subject to confounding—in such a way that, on repeated testing, apparently mixed patterns of significant and non-significant results may be obtained.

(2) To also demonstrate, more generally, that where a confounding factor is suspected that could influence the outcome of a hypothesis test, it is possible by resampling-based simulation to chart the *sensitivity* of the planned test to various levels of the potential confounding. If this process is applied systematically, it would be possible to construct a curve, analogous to a power curve, showing varying degrees of the potential for confounding-based error, under specified conditions.

Regarding (2), it is acknowledged that if a confounding factor is suspected, it would be even better (as noted earlier) to re-design the experiment, if possible, to avoid the confounding. But if this is not feasible—or if the experiment has already occurred—Objective 2 could provide an indicator of whether the confounding bias is likely to affect the outcome.

Research Model

The following observations provide the basis for developing the planned demonstrations. When a

parametric hypothesis test is conducted, it is presupposed that there is a well-defined null distribution, whose central value is meaningful and accurately modeled. This might appear just to be a restatement of the null hypothesis concept; but as shown in the following example, that is not quite the case: suppose inspectors are sampling from a warehouse of presumed "9-volt" batteries. They take a large sample, and test for a mean sample voltage of 9 volts. *If there is no confounding*, then obviously the null distribution should be centered on 9 volts, and H_0 is rejected if the sample mean is sufficiently far above or below that value.

But consider the effect on the paradigm if a confounding factor exists. Suppose that the measuring instrument is defective—the voltmeter is biased to give lower-than-accurate readings. If this fact were known, then the experimenters would obviously replace the instrument, if possible. But if they proceed in ignorance, then the null distribution used in their model does *not* accurately reflect the voltages-*as-measured* that would actually be expected given their intended null hypothesis: Taking a sample of batteries from a population with a mean voltage of 9 volts, but planning to read the voltages with a tool biased downwards, the true center of the expected null distribution of readings is some value less than 9 volts.

Extend the case to consider that (a) there is no replacement voltmeter available, and (b) the presence and/or degree of bias in the instrument is not known. This scenario would motivate discovering some method to assess how sensitive the planned, conventional test would be to the confounding, if indeed it is present. That is, how much confounding effect could be tolerated without making a difference in the outcome of the conventional test, and at what point does the effect make a difference? This paper illustrates a means to answer such questions.

Methods

As noted above, confounding denotes the presence of a variable that is "is associated with both the 'exposure' or independent variable and with the 'outcome' or dependent variable under study" (Civetta 1999). But as Pearl observes (2001), to control for confounding, some assumptions must also be made about the *causal* relationships in the problem. Causal assumptions also apply for the present method: a confounding relationship is identified that is seen as causally plausible (though its exact strength is not known); measures are then developed to assess the sensitivity of a conventional hypothesis test to bias, induced by that confounding.

The proposed approach has similarities to simulation-based calculations for the power of a hypothesis test, such as applied by Vickers (2001) in evaluating a medical research model. Power addresses the probability that a test will not fail to recognize an actual difference in a parameter value (say, the mean) from the value assumed in H_0. One can evaluate a test's power by supposing the true parameter to have some known, different value from the H_0 hypothesis, and generating a simulated sampling distribution from that alternative population. One observes empirically the proportion of simulated samples for which the test would—correctly—identify a significant difference in the parameter from H_0; this approximates the power.

The simulation below also posits a difference between the *actual* tested population (the "corrected population") and the presumed population that is implicit in H_0. One generates a simulated sampling distribution from the corrected population, then one observes empirically: what range of values in the simulated sampling distribution (a) fall in value ranges that conventionally would be considered in the critical regions, but (b) are, with respect to the corrected population, *non*-significant? The area under this part of the distribution curve gives some measure of the sensitivity of the conventional tests to bias, if in fact the corrected population reflects a real causal relationship at play.

The case for illustration is modeled on the published findings found in McCarthy, Schuster, Zehr, and McDougal (1999), who perceived that they found a significant increase in the critical thinking (CT) competencies as measured by the instrument CCTDI when comparing the scores of seniors to those of sophomores. The conventional null distribution for their test modeled the hypothesis of "no difference" between the two years' results. No distinction was acknowledged between (a) "H_0: no difference in (real) CT competencies" and (b) "H_0: no difference in CT-scores-as-measured-by-the-instrument". If there is some reason that the score might reflect, as well, an extraneous variable, then the two hypotheses are not truly equivalent.

As mentioned in the Background section, the author has observed that McCarthy's chosen CT-assessment instrument may be non-responsive, in the sense of eliciting inconsistent answers to the same questions from the same people during test and retest. However, in the midst of that apparent randomness of before-and-after answers, some "currents" of order, not exactly trends, were found: Some responders (who might be called "good day'ers") demonstrated a slight, but uneven tendency to give more positive answers on the retest opportunity than on the original test. Other responders

(call them "bad day'ers") demonstrated the opposite, slight propensity to give more negative responses. The dispersion of re-answers for both groups was wide, and did not preclude some "good day'ers" giving lower or unchanged responses, or "bad day'ers" giving higher or unchanged responses. Still, a small probabilistic tendency was perceivable.

These "microtrends" in the data, based on a "good-day/bad-day" effect (or other, unknown cause), would add to the variance of the true null distribution for "no difference" in a manner not accounted for in the conventional hypothesis test. We do not know if this effect applied to the samples collected by McCarthy *et al*; nor how powerful the effect would be if it is present, nonetheless, we can now check how sensitive their test might be to error if the possible effect is ignored.

The simulation design echoes the sampling scenario employed by McCarthy *et al*. Virtual samples were created for 156 sophomores and 85 seniors, as in McCarthy's paper. Each sampled student obtained a randomly generated score, constructed analogously to calculating a score on the CCTDI: that is, (ignoring subscales) the maximum score is 420, based on the weighted distribution of all answers to individual questions, each having a value from 1-6. The confounding factor "good or bad day" was added in this fashion: each "student", independently, was randomly determined to be a "good day'er" or a "bad day'er", and the former exhibited a *slight* stochastic tendency to shift their answer distributions towards the higher end; the latter exhibited a *slight* stochastic tendency to shift answer distributions towards the lower end (the Excel encoding for these "tendencies" can be examined in the attached data files; also see the Appendix for explanations of the fields in the worksheet).

A computer model that embodies the above tendencies was run to generate 5000 independent samples. After each run (representing one sample), the mean simulated score for the seniors was compared with the mean score for the sophomores. The percentage change in means between the two groups was interpreted as the test statistic for the sample. The relative frequency distribution for the test statistic estimates the full sampling distribution for a population that has the same specified characteristics. The results reported by McCarthy were assessed in the light of that distribution. The experiment was then repeated (a) assuming no presence of confounding, and (b) a small confounding effect—but weaker than in the first experiment.

Selection of the Simulated number of samples R

The simulations described were generated by resampling from a distribution interpreted as reflecting the confounder-modified population, from which an unsuspecting researcher might draw a sample. Random samples of a given size were repeatedly generated from the modeled population, so the test statistic could be calculated for each sample, and the overall distribution of the statistic observed. The question had to be addressed of how many resamples are required so that this distribution approximates a full sampling distribution of the statistic from the population?

There is no definitive answer to this question. The number of resamples R = 1000 is a common choice, employed for example with little or no discussion in Vickers (2001), Blank, Seiter & Bruce (2001), and Sormani *et al.* (1999). Yet, as Davison and Hinkley (1997) observe, "In any simulation-based test, relatively few samples could be used if it quickly became clear that *p* was so large as not to be regarded as evidence against H_0" . In their own text, a variety of R's, starting as low as 49, are used in different examples, although, as in other books, a value of 1000 (or sometimes 999) is used most frequently. The key point is to simulate enough samples to estimate the shape and location especially of the distribution tail. There is little computational cost in choosing R greater than 1000, to further smooth and remove gaps in the curve, and generally resampling with large R's "introduces little variation" (Hesterberg *et al.* 2003). In fact, Hesterberg refers to an application by the U.S. local telephone company Verizon, in which they use an R of 500,000. Based on these considerations, the simulation below employs an R of 5000, to smoothly model the curve to be expected from a sampling distribution.

(*Note: Details on interpreting the attached data set are found in the Appendix, following the references.*)

Findings

The change in students' mean scores reported as significant in McCarthy *et al* was from 315.48 to 325.95, an increase of 3.32%. But based on the simulated sampling distribution—as adjusted for suspected confounding in the first experiment—it was found that increases of that magnitude or greater occurred by chance about 5.2% of the time. (See Figure 1, solid line.) That is, the reported increase in scores is not significant on the adjusted model. *If* this model captures a true effect related to the measurement instrument employed to capture the test statistic, then it could explain why it is not unusual for a few studies to find

improvements in scores, independently of whether a true improvement in CT skill has occurred.

If the modeled confounding effect is modified, to be less pronounced than in the first experiment, then the adjusted model for the null distribution is closer to the conventional H_0. In that case, the presence of confounding does not significantly alter the results from those published by McCarthy *et al*; namely, α appears significantly small. This suggests that the conventional test results would not be affected by confounding, for effects that are less than modeled in the first experiment.

Discussion

It is common in scientific studies to test whether the introduction of some treatment makes a difference (has an effect) in some domain. In the above case, the domain is the attempt by nurse educators to impart critical thinking skills to their students. If students are evaluated at two stages of a program, then presumably the education occurring between the stages is the "treatment". The test statistic can be based on an increase in students' Critical Thinking (CT) scores between the program's two respective stages. But this requires that a reliable measurement tool be available to obtain the scores in both periods (also note that for uncontrolled studies, other influences may also be influencing the results).

As described earlier in this paper, the extensive literature on attempts to measure CT-development creates a disturbing impression that changes in critical thinking may be nearly impossible to measure or to accomplish through education. Many studies have had "mixed" results, and the minority of studies having clearer results are nearly balanced between positive or negative findings. Rather than make conclusions about CT training, one could alternatively question the assessment tools that have been utilized. This paper's Case has referred to a specific instrument (the CCTDI), but Tucker (1996) and Wilson (2000) clearly indicate that the measurement problems indicated might be more global.

While these historical issues provide the context for this paper, its main interest is statistical. If a flawed instrument *is incapable* of measuring a treatment effect, we would not expect to measure "significant differences" between groups with the tool, beyond the occasional case of sampling error. Yet in the Case presented, we see one study which does purport to find a significant difference, using the tool.

This effect could be explained if the test process is subject to confounding. A working hypothesis is that, while, in

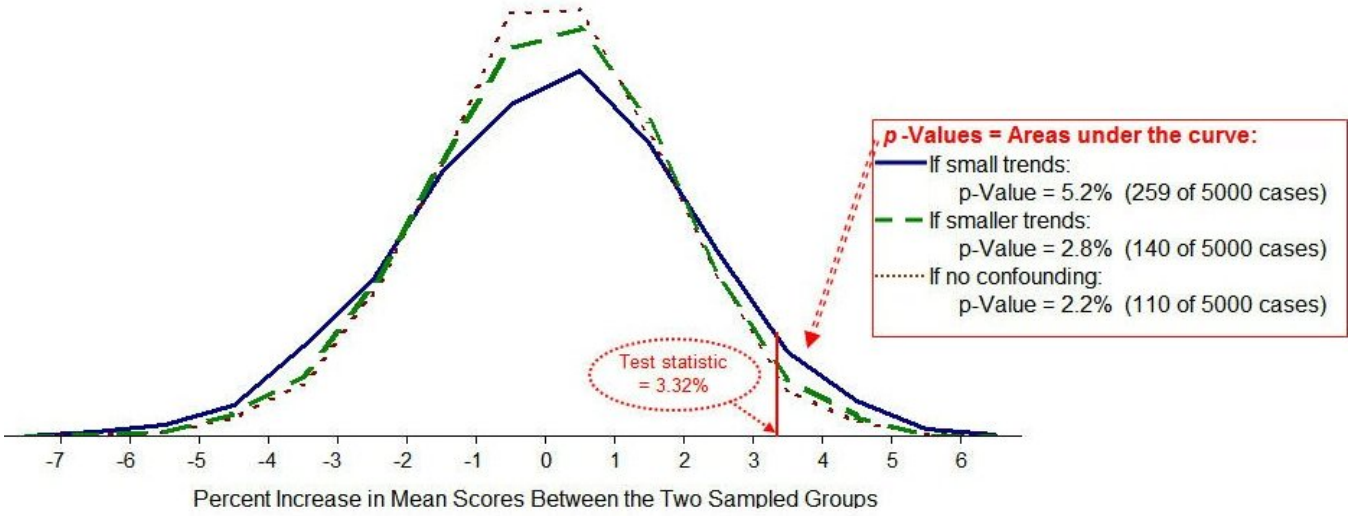

Figure 1. Three Simulated Sampling Distributions.

general, students' scores are nearly random, there are small tendencies for individual students to answer with an upward-score bias or downward-score bias on any given day. This creates the effect that the null-distribution-as-measured (i.e. the pattern of expected sampling statistics if the original null is true) is not exactly identical to the null-distribution-as-modeled by a conventional hypothesis test. Figure 1 shows the implications: A sampling result that appears unlikely on the conventional null hypothesis (p-value = 0.022) is in fact, non significant if the confounding factor is reflected in the model (p-value = 0.052).

Figure 1 also shows the basis for constructing a table or curve, analogous to a power curve: If a test is subject to confounding, it is plausible that one will not know the exact strength of the effect. Examination of the Excel-based model of the effect would show a distinction between the "small" effect and the "smaller" effect. But it would require causal knowledge of the context to know which model, if either, best matches the real strength and dispersion of the confounding effect. Nonetheless, we can perform a what-if analysis of which types of scenarios could plausibly bias a conventional test's results, and which scenarios would have trivial effects, if any.

Conclusions

Examining this Case has led to two basic findings—one specific, and one general. The first concerns the requirement for a reliable measurement tool in conducting a hypothesis test. Failing that, the resulting test statistic could be unpredictable, and results could be influenced unduly by small, but unknown confounding effects.

The second finding is, essentially, a proof of concept. It is illustrated how resampling techniques can be used to model the possible impact of confounding effects on an (assumed) null hypothesis. The suspected impact of the confounder can be modeled to simulate the sampling-distribution-as-measured that would follow if (a) the original null hypothesis is true, and (b) the strength and dispersion of the confounding effects are as modeled. Re-running the simulation with varied degrees of strength for the confounder, provides an indicator of how strong the effect must be before it becomes a concern, for interpreting the conventional hypothesis test.

When designing a *new* test, e.g. for assessment of applying a treatment, it would of course be preferable to identify the possible causes of confounding in advance, and to control for them in various appropriate ways—such as improving the test instrument, designing controlled experiments, reconsidering sample size, and so on. But if not all these options are realistic, or if a suspect test has already occurred, then the proposed method provides one way to guide our level of confidence in the results.

The author proposes that future work be undertaken to generalize this second finding, in particular, and to elucidate some more detailed principles and procedures for how it should be conducted.

Correspondence: Bill.Goodman@uoit.ca

REFERENCES

Adams, B.L. 1999. Nursing education for critical thinking: An integrative review. *Journal of Nursing Education*, 38(3), 111-119.

Adams, M.H., Stover, L.M., & Whitlow, J.F. 1999. A longitudinal evaluation of baccalaureate nursing students' critical thinking abilities. *Journal of Nursing Education*, 38(3), 139-141.

American Association of Colleges of Nursing 1998. *The Essentials of Baccalaureate Education for Professional Nursing Practice.*

Angel, B.F., Duffey, M., & Belyea, M. 2000. An evidence-based project for evaluating strategies to improve knowledge acquisition and critical- thinking performance in nursing students. *Journal of Nursing Education*, 39(5), 219-228.

Argyris, C. 1982. *Reasoning, learning, and action: Individual and organizational.* San Francisco: Jossey-Bass.

Beckie, T.M., Lowry, L.W., & Barnett, S. 2001. Assessing critical thinking in baccalaureate nursing students: A longitudinal study. *Holistic Nursing Practice*, 15(3), 18-26.

Berger, M.C. 1984. Clinical thinking ability and nursing students. *Journal of Nursing Education*, 23(7), 306-308.

Blank, S., Seiter, C., & Bruce, P. 2001. *Resampling Stats in Excel. Version 2.*. Arlington,VA: Resampling Stats Inc.

Bowles, K. 2000. The relationship of critical-thinking skills and the clinical judgment skills of baccalaureate nursing students. *Journal of Nursing Education*, 39(8), 373-376.

Box, G.E.P., Hunter, J.S., & Hunter, W.G. 2005. *Statistics for experimenters: Design, innovation, and discovery.* Second Edition. Hoboken, N.J.: John Wiley and Sons.

Brown, J.M., Alverson, E.M., & Pepa, C.A. 2001. The influence of a baccalaureate program on traditional, RN-BSN and accelerated students' critical thinking abilities. *Holistic Nursing Practice*, 15(3), 4-8.

Brown, Race, & Smith 1996. An assessment manifesto. Published on the *Deliberations* website, maintained by London Guildhall University. http://www.city.londonmet.ac.uk/deliberations/assessment/manifest.html.

Cassel, J.F. & Congleton, R.J. 1993. *Critical thinking: An annotated bibliography.* London: Scarecrow Press.

Civetta, J.M. 1999. Statistics, the literature, hospital data and patient profiles: A survival guide. *The Internet Journal of Anesthesiology*, 3(4), http://www.ispub.com/journals/IJA/Vol3N4/literat

ure.htm.

College of Nurses of Ontario 1999. *Entry to practice competencies for Ontario registered practical nurses.* Toronto, Ontario: College of Nurses of Ontario (website: http://www.cno.org/docs/reg/41042_EntryPracRPN.pdf)

Colucciello, M.L. 1997. Critical thinking skills and dispositions of baccalaureate nursing students--A conceptual model for evaluation. *Journal of Professional Nursing*, 13(4), 236-245.

Davison, A.C. & Hinkley, D.V. 1997. *Bootstrap methods and their application.*. Cambridge, U.K.: Cambridge University Press.

Dewey, J. 1959. *Dewey on education: Selections*. Edited by M.S. Dworkin. NY: Teachers College Press.

Edwards, S. 2003. Critical thinking at the bedside: A practical perspective. *British Journal of Nursing*, 12(19), 1142-1149.

Ennis, R.H. 1996. *Critical Thinking.* Upper Saddle River, NJ: Prentice Hall.

Facione, N.C. & Facione, P.A. 1997. *Critical thinking assessment in nursing education programs: An aggregated data analysis. Millbrae, CA: The California Academic Press.*

Facione, P.A. 1990. *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction.* ERIC ED 315 423, American Philosophical Association.

Facione, P.A., Facione, N.C., & Giancarlo, C.A. 1997. *The motivation to think in working and learning.* Millbrae, CA: Insight Assessment (website: http://www.insightassessment.com/pdf_files/Motivatn_Thnk_Wrk_Lrn_19 97.PDF)

Facione, P.A., Facione, N.C., & Giancarlo, C.A. 2000. The disposition toward critical thinking: Its character, measurement, and relation to critical thinking skill. *Informal Logic*, 20(1), 61-84.

Facione, P.A., Facione, N.C., & Giancarlo, C.A. 2001. *California critical thinking disposition inventory: Inventory manual.* Millbrae, CA: The California Academic Press.

Giancarlo, C.A. & Facione, P.A. 2001. A look across four years at the disposition toward critical thinking among undergraduate students. *The Journal of General Education*, 50(1), 29-55.

Goodman, WM 2004. Can Critical Thinking Skills Be Taught? A Study on the Use of Interactive Web-Centric Technologies to Enhance These Skills among Students in a BScN Program. Presented at the 22nd Annual International Nursing Computer and Technology Conference sponsored by Rutgers College of Nursing Center for Professional Development. Arlington, Virginia.

Gordon, J.M. 2000. Congruency in defining critical thinking by nurse educators and non-nurse scholars. *Journal of Nursing Education*, 39(8), 340-351.

Gross, Y.T., Takazawa, E.S., & Rose,C.L. 1987. Critical thinking and nursing education. *Journal of Nursing Education*, 26(8), 317-323.

Habicht, J.-P., Mason, J.B., and Tabatabai, H. 1984. Basic concepts for the design of evaluation during programme implementation. Chapter 1 in Sahn, D.E., Lockwood, R., and Scrimshaw, N.S., Editors, *Methods for the Evaluation of the Impact of Food and Nutrition Programmes*. Tokyo: United Nations University Press.

Hair, Anderson, Tatham, & Black 1998. *Multivariate analysis*. Fifth edition. Upper Saddle River, NJ: Prentice Hall.

Hesterberg, T., Monaghan, S., Moore, D.S., Clipson, A., & Epstein, R. 2003. *Bootstrap methods and permutation tests: Companion chapter 18 to the practice of business statistics.* New York: W.H. Freeman and Company.

Magnussen, L., Ishida, D., & Itano, J. 2000. The impact of the use of inquiry-based learning as a teaching methodology on the development of critical thinking. *Journal of Nursing Education*, 39(8), 360-364.

Martin, C. 2002. The theory of critical thinking of nursing. *Nursing Education Perspectives*, 23(5), 243-247.

May, B.A., Edell, V., Butell, S., Doughty, J., & Langford, C. 1999. Critical thinking and clinical competence: A study of their relationship in BSN seniors. *Journal of Nursing Education*, 38(3), 100-110.

McCarthy, P., Schuster, P., Zehr, P., & McDougal, D. 1999. Evaluation of critical thinking in a baccalaureate nursing program. *Journal of Nursing Education*, 38(3), 142-144.

National League for Nursing 2003. *Critical thinking in clinical nursing practice/RN examination.* Web-published bulletin posted at http://www.nln.org/testprods/pdf/CTInfobulletin.pdf

National League of Nursing 2004. Innovation in nursing education: A call to reform. *Nursing Education Perspectives*. 25(1), 47-49.

Norris, S.P., Editor 1992. *The generalizability of critical thinking: Multiple perspectives on an educational ideal.* NY: Teachers College, Columbia University.

Ohio University 2002. School of Nursing 2001-2002

student learning outcomes assessment report. Web-published by the office of Provost, Ohio University at www.ohiou.edu/provost/SLOA2001_2002/nursing2002doc

Pearl, J. 1998. Why there is no statistical test for confounding, why many think there is, and why they are almost right. *Technical Report R-256,* Cognitive Systems Laboratory, UCLA. July. (Draft Copy).

Pearl, J. 2001. Causal inference in the health sciences: A conceptual introduction. *Health Services & Outcomes Research Methodology*, 2, 189-220.

Pepa, C.A., Brown, J.M., & Alverson, E.M. 1997. A comparison of critical thinking abilities between accelerated and traditional baccalaureate nursing students. *Journal of Nursing Education*, 36(1), 46-48.

Saltelli, A. 2002. Sensitivity Analysis for Importance Assessment. *Risk Analysis*, 22(3), 579-590.

Scheffer, B.K., & Rubenfeld, M.G. 2000. A consensus statement on critical thinking in nursing. *Journal of Nursing Education*, 39(8), 352-359.

Schon, D.A. 1987. *Educating the reflective practitioner.* San Francisco, London: Jossey-Bass.

Sormani, M.P., Molyneux, P.D., Gasperini, C., Barkhof, F., Yousry, T.A., Miller, D.H., & Filippi, M. 1999. Statistical power of MRI monitored trials in multiple sclerosis: New data and comparison with previous results. *Journal of Neurology, Neurosurgery, and Psychiatry*, 66(April), 465-469.

Spelic, S.S., Parsons, M., Hercinger, M., Andrews, A., Parks, J., & Norris, J. 2001. Evaluation of critical thinking outcomes of a BSN program. *Holistic Nursing Practice*, 15(3), 27-34.

Tawari, A. 2004. [Abstract of a paper under review, obtained through personal correspondence with the author.]

Tucker, R.W. 1996. Less than critical thinking. *Assessment and Accountability Forum*, 6 (numbers 3 and 4).

Vickers, A.J. 2001. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: A simulation study. *BMC Medical Research Methodology*, 1(6).

Walsh, C.M., & Hardy, R.C. 1999. Dispositional differences in critical thinking related to gender and academic major. *Journal of Nursing Education*, 38(4), 149-155.

Wilson, R.W. 2000. Evaluative properties of critical thinking tests: Change scores from students in physical therapy and other health care professions. *Journal of Physical Therapy Education*, 14(2), 27-31.

Appendix: Notes on the Attached Data Sets

All data are in the file **goodman.xls**.

Tab: **SamplingModels**

This worksheet includes three simulations that model the test actually conducted and reported by McCarthy *et al*. For the assumption of "no confounding effect present," **Column G** contains simulated scores for the 156 students in Group 1, and **Column AD** contains simulated scores for the 85 students in Group 2. The mean scores for both groups are recorded in cells **W6** and **W13**, respectively; and the difference in means is found in cell **W22**. By clicking the *F9 key* in Excel, all inputs change randomly within the constraints of the model, to produce a second simulated sample, with a different result in **W22.** If F9 is clicked multiple times, and all samples' results are recorded, a frequency distribution for all sample results can be constructed that approximates a sampling distribution for the sample statistic.

A similar structure is used for the other two simulations. For the assumption of "a small confounding effect is present," **Column O** contains simulated scores for the 156 students in Group 1, and **Column AL** contains simulated scores for the 85 students in Group 2. The mean scores for the groups are recorded in cells **W8** and **W15**, respectively; and the difference in means is found in cell **W24**. For the middle case of a weaker confounding effect, **Columns V** and **AS** are used, respectively, for the students' scores in the two groups, and cells **W10**, **W17**, and **W26** are used for the group means and their difference. Again, clicking the *F9 key* changes all inputs randomly within the constraints of the model, to produce different results.

Students are encouraged to explore more details of the model by examining the formulas within the cells. Random numbers are generated in the "Work Areas" of the worksheet, in the first six columns of each the areas labeled (a) and (d). For Groups 1 and 2 of the "no confounding effect" model, the relative sizes of those six random numbers directly determine, for each "student" (i.e. for each row in the model), the relative proportions of his/her answers (for 70 questions) that are 1's, 2's, 3's, 4's, 5's, or 6's. From this the student's simulated score can be calculated. But for the groups with a "small confounding effect" or a "smaller confounding effect", the relative proportions of a student's answers (for 70 questions) that are 1's to 6's is determined by adjusting what the student would get on the "no confounding" model by a randomly determined "good day/bad day" variable (in **Column H**) together with a model for the impact of the effect.

Tab: 5000Results

Based on the model just described, the results of 5000 simulated samples were taken, for each of the two Groups, with respect to the three possible assumptions for the confounding effect (i.e. no effect, small effect, or smaller effect). In practice, the author used the add-in software *Resampling Stats for Excel*, and the results for all 5000 tries are summarized in this second spreadsheet. However, one could replicate the results (with random variation) with alternative software, or (if there is enough time) by many clicks of *F9*, and recording the results manually.

The columns for relative difference in means were added afterwards. The distributions of values in these three columns provide the basis for Figure 1 in the text.