Critical Review

CAT Scan: A Critical Review of the *Critical-Thinking Assessment Test*

KEVIN POSSIN

Professor Emeritus, Philosophy
Winona State University
The Critical Thinking Lab
1012 Calle Dorthia
Santa Fe, NM 87506
USA
kpossin@winona.edu

Abstract: The CAT is entirely dedicated to assessing the critical-thinking skills involved in scientific reasoning and practical problem solving. While the test is found to have reasonable content validity, various issues with its prompts are discussed, along with significant issues with its scoring. The CAT's recommended use as a "model" for curricular changes, called CAT Apps, is criticized as "teaching to the test."

Résumé: Le CAT porte sur l'évaluation des compétences en pensée critique impliquées dans le raisonnement scientifique et dans la résolution de problèmes pratiques. Bien que le test ait une « content validity » raisonnable, divers problèmes liés à ces brefs passages de texte sont discutés, ainsi que des problèmes importants liés à sa notation. L'utilisation recommandée par le CAT comme «modèle» pour les changements curriculaires, appelée applications CAT, est critiquée comme un enseignement axé exclusivement sur l'examen.

Keywords: CAT Apps, critical thinking, critical-thinking assessment tests, problem-solving skills, scientific-reasoning skills, validity: content, construct, and criterion

1. Introduction

There are numerous critical-thinking assessment tests commercially available, all with decades of use and research in support of them and critical reviews questioning them. Most of these tests are for the

purposes of assessing a wide array of *general* critical-thinking skills, determined by a general definition of critical thinking, such as "a process, the goal of which is to make reasonable decisions about what to believe and what to do" (Ennis 1996) or "the practice of requiring, assessing, and giving cogent reasons for one's beliefs, values, and actions" (Possin 2002).

This account of critical thinking can roughly be analyzed into the following component skills or competencies:

- Identifying *reasons* or *arguments*
- *Analyzing* or dissecting arguments into *premises*, *conclusions*, and *subconclusions* (explicit and implicit)
- Taxonomizing arguments as *deductive* or *inductive*
- Assessing the *cogency* of arguments, *relative to their type*, in terms of the truth or *acceptability* of their premises and the *relevance* of their premises as indicating the truth or probable truth of their conclusions
- Identifying *formal* and *informal* fallacies—in essence, popular ways of *failing* these cogency conditions
- Critically reviewing *definitions* and *analyzing concepts*
- Assembling these competencies so as to select and argue for rational positions on a diversity of issues, critically reviewing competing positions and their arguments, all in a cogent and intellectually honest manner (Possin 2002, 2008).

Sometimes, however, one is interested in assessing the status and development of a certain *subset* of these critical-thinking skills, for example, those involved in inductive and scientific reasoning and practical problem solving and decision making. Perhaps this is for the purpose of assessing how effective certain STEM or scientific-reasoning courses are at enhancing the focused critical-thinking skills they are designed to promote. The first rule of assessment is, in any case, "pick the right tool for the job" (Hatcher and Possin 2020). So, how well do prominently marketed critical-thinking assessment instruments test for the critical-thinking skills involved in scientific reasoning, decision making, and problem solving?

In this review, I want to take a close look at a fairly recent, and more specialized, entry to the field of critical-thinking tests, the *Critical-Thinking Assessment Test* (CAT), in this capacity. Before I begin that

© Kevin E. Possin.

project, however, I would like to very briefly mention the extent to which some of the competing *general* critical-thinking tests might serve our more focused purpose, since some, for example, the *Ennis-Weir Critical Thinking Essay Test*, do *not* at all (despite being modestly good at assessing some critical-thinking skills as manifested in argumentative writing [Possin 2008]).

- The *Collegiate Learning Assessment* (CLA+) *Test* has an hour-long portion dedicated to a performance task that involves the use of supplied documents to write a position paper arguing for one's decision (and against competing positions) on how to solve a prescribed problem. This portion of the CLA+ is quite good, judging from sample performance tasks presented and scored on the CLA+ website over the years; the main problem is that its *scoring* can be manipulated in ways counter to critical thinking (Possin 2013). Furthermore, the CLA+'s other half-hour portion contains only 10 multiple-choice items dedicated to "scientific and quantitative reasoning," which is too few for our purpose of assessing scientific-reasoning skills.
- The *California Critical Thinking Skills Test* has only 13 of its 40 items (32%) dedicated to inductive reasoning, which it characterizes as making decisions under uncertainty to reasonable and probable conclusions, based on, for example, experience, statistical analysis, testimony, analogies, and patterns of events.
- The *Cornell Critical Thinking Test Level Z* has only 17 of its 52 items (33%) dedicated to identifying premises and conclusions of inductive arguments.
- The *Watson-Glaser Critical Thinking Appraisal* has 23 of its 40 items (56%) dedicated to inductive reasoning (in its Making Inferences and Interpreting Arguments sections) and to defeasibility reasoning (in its Evaluating Arguments section). Unfortunately, the directions for these sections are so confusing that it is unclear whether one is being tested on inductive reasoning or *deductive* reasoning (Possin 2014), which is why Possin's subtitle is, "The More You Know, the Lower Your Score."

- The *Halpern Critical Thinking Assessment Test's* Decision Making and Problem Solving section makes up 31% of the test, with the sections Hypothesis Testing and Judging Likelihood and Uncertainty making up 36%, for a total of 67%. This would make the HCTA a worthy candidate for the job if it were not plagued by serious problems with, for example, erroneous scoring keys, a highish cost, and a less than user-friendly online platform for administrators, scorers, and test takers (Possin 2013).

2. The CAT

In comparison, the *Critical-Thinking Assessment Test* would *seem* to be exactly what we're looking for—*entirely* dedicated to testing one's scientific-reasoning and problem-solving skills, with no argument analysis, no identification of fallacies (formal or informal), and *no* testing of formal deductive "logical reasoning" skills (a fact seemingly celebrated in [Stein, Haynes, and Redding 2006, p. 291]).

The CAT is a short-answer essay exam, involving 15 questions that can be completed within an hour, although it is *not* a timed exam (to its credit! critical thinking is "slow thinking," not "fast thinking" [Kahneman 2011]). Questions in Part I revolve around various realistic scenarios. We are asked to evaluate hypotheses, provide alternative explanations of data, and suggest additional information that would help us further evaluate those hypotheses. Part II presents us with a real-world problem to solve: deciding which piece of equipment would best suit a family taking an extensive camping trip. To make our decision, we must select relevant sources of information from among documents provided in the exam and use that information to make the *best* decision from among the alternatives offered, giving reasons for our choice. We are also asked to discuss how changes in the scenario's circumstances would/should affect our proposed solution.

I would say that, by good fortune, the CAT *per se* quite successfully tests for the subset of "core critical-thinking skills" it was designed to measure and that these *are indeed* core critical-thinking skills:

Evaluating Information

Separate factual information from inferences.

Interpret numerical relationships in graphs.
Understand the limitations of correlational data.
Evaluate evidence and identify inappropriate conclusions.

Creative Thinking

Identify alternative interpretations for data or observations.
Identify new information that might support or contradict a hypothesis.
Explain how new information can change a problem.

Learning and Problem Solving

Separate relevant from irrelevant information.
Integrate information to solve problems.
Learn and apply new information.
Use mathematical skills to solve real-world problems.

Communication

Communicate ideas effectively. (Stein and Haynes 2011, p. 45)

In discussing the development of the CAT, Stein and Haynes describe how, in 2000, a team of interdisciplinary faculty from Tennessee Tech University came to a consensus on this list of what they believed to be important critical-thinking skills. Those, along with other faculty, later guided the drafting of the CAT questions (which mirror the skills listed above, often verbatim) and how possible answers to them should be scored by faculty eventually administering the test. It is little wonder, then, that 90% of those 69 interdisciplinary faculty from six universities later rated the resultant questions on the CAT as valid measures of critical-thinking skills (with the exception of one item involving an elementary math calculation, which was rated valid by only 80%) (Stein and Haynes 2011, p. 45). The CAT is thereby said to have high *face validity*, which simply means that it *seems* to the raters to be valid. Having such high face validity in the eyes of faculty (and test takers) is important, as evidence that the test will be given (and taken) seriously. But how good is it as evidence of a test's *content validity*, defined as being an *accurate* measure of *actual* critical-thinking skills (just like an accurate gas gauge measures the actual amount of fuel in one's tank)? High face validity is only as good as the assumed expertise of the judges. If that assumption is false, judgments of high face validity are as indicative of content validity as President Trump's opinion regarding the high face validity of Vladimir Putin's claim that Russia did *nothing* to influence the 2016 election based on how "strongly" Putin made that claim.

© Kevin E. Possin.

As famously illustrated by C. H. Lawshe (1975), "In achievement testing we normally use 'subject matter experts' to define the … 'content domain.' [….] If the subject matter experts are generally perceived as true experts, then it is unlikely that there is a higher authority to challenge the purported content validity of the test" (p. 565). The question is, then, who are the experts? A group of faculty from across the disciplines, who claim to teach some critical thinking? No. Paul, Elder, and Bartell (1997) surveyed numerous interdisciplinary faculty and found that, while they all thought critical thinking is essential and thought *their* curricula enhanced students' critical-thinking skills, these faculty were either embarrassingly ignorant of the core elements of critical thinking or woefully mistaken about them, often thinking critical thinking and truth are purely *subjective*. Unfortunately, Kruger and Dunning (1999) found that *competence* in logical reasoning is *inversely related to one's confidence*. Stein, Haynes, and Redding even *admit* that "Most faculty have little training in developing classroom assessment that promotes the development of critical thinking skills" (2016, p. 2). So, who has training in critical thinking; who are the experts? Those who *specialize* and do *research* in critical thinking; those who teach *dedicated* courses in critical thinking; those who understand critical-thinking skills in their *generic* and *transferable* form, not just some *discipline-specific* instantiation of them—*they* are the experts.

So far, then, the CAT would seem to have meager evidence for its content validity. How about its *construct validity*? "Learning-sciences and assessment experts" were called upon "for the sake of construct validity … the questions [and associated scoring process] had to have construct validity from the perspective of contemporary theory in the cognitive/learning sciences" (Stein and Haynes 2011, p. 45). The question now becomes, are those experts in the *learning sciences* and *assessment* experts in critical thinking? And the answer is, not really. Construct validity, by definition, begins with the philosophical analysis of an abstract concept (i.e., the "construct" one wants to measure) into its essential elements, each of which is then "operationally defined" via observational indicators, which become candidates or templates for test items. In our case, the "construct" is critical thinking, to be analyzed into its component critical-thinking skills to be tested for by means of the items in the assessment test. Without the assistance of *real expertise in critical thinking*, however, construct validity is an

© Kevin E. Possin. *Informal Logic*, Vol. 40, No. 3 (2020), pp. 489-508.

unanchored, question-begging practice, like buying numerous copies of *your own* newspaper to assure yourself that *your own* story is correct (to paraphrase Wittgenstein [1953, §§265]).

Stein, Haynes, and Redding (2016, p. 6) claim *criterion validity* for the CAT, by citing significant CAT gains shown with National Science Foundation projects "designed" to enhance critical-thinking skills. However, they also admit that out of the approximately 40 such projects, only about *half* showed such gains. At the risk of rendering this claim to the CAT's validity unfalsifiable, then, we must admit that it is a toss-up whether, in that *other* half, those projects failed *or the CAT did*. However, getting significant gains on the CAT after *actual* instruction in critical thinking would *not* be evidence of *criterion validity*, which I will discuss in a moment. *Technically*, it would at best be evidence for the CAT's *content validity*.

> [O]bservations of the sensitivity of the [CAT] to the effects of high impact practices in education designed to impact these [critical-thinking] skills have been used to support the validity of the instrument (Stein et al., 2007). The test has been sensitive to the effects of high impact educational practices in both formal and informal settings that span a semester or less (Alvarez, Taylor and Rauseo, 2015; Carson, 2015; Gasper, Minchella, Weaver, Csonka, and Gardner, 2012; Gottesman and Hoskin, 2013; Rowe et al., 2015). (Haynes et al. 2015, p. 40)

Unfortunately, (Stein, Haynes, Redding, Ennis, and Cecil 2007) contains no such evidence of the CAT's content validity. Most results in (Alvarez et al. 2015) are not statistically significant, and no effect sizes are provided. And no results are given in (Gasper et al. 2012). But then, neither of these latter two cases involved significant critical-thinking *interventions* to even test the effects *of*. In the other three instances mentioned by Haynes et al., statistically and practically significant gains were indicated by the results of post-course CAT scores, and the courses had significant amounts of critical-thinking content. I will comment further on these cases later when I discuss "CAT Apps." At this time, I mention them as acknowledged evidence of the CAT's content validity.

Evidence of *criterion validity* would be, by definition, a matter of demonstrating a correlation of gains on the CAT with gains on *another valid critical-thinking assessment indicator*. (Criterion validity is like verifying the accuracy of your gauge by confirming that its readings match those of another gauge that is independently known to be

accurate.) Indeed there is some evidence of this, as gains on the CAT were found to be correlated with gains on the *California Critical Thinking Skills Test* (R = .645) and the *Collegiate Assessment of Academic Proficiency-Critical Thinking Test* (R = .691) (Center for Assessment and Improvement of Learning 2016). However, the set of critical-thinking skills that the CAT assesses overlaps so little with the critical-thinking skills assessed by the other two tests that I am not confident that these correlations are very evidential.

Despite this paucity of evidence for the overall validity of the CAT, I am delighted to say that the CAT is more or less spot on in testing for the important subset of critical-thinking skills it claims to assess—the *test's* content validity is a most happy coincidence, and I will make only a few suggestions as to how to improve it. This good fortune has its limits, however. While I have praises for the *test per se*, I have some serious concerns about the CAT's *scoring guide* or *answer key* that faculty scorers are trained to use—concerns that I believe reduce the construct validity of the assessment *process* when using the CAT. Think of it this way: a gauge can be quite accurate and yet be regularly *misread* by those using it; and sometimes the gauge might even *invite* this misreading. Let me explain.

Items 1-4 of the CAT concern the likes of the following scenario: Over the past decade, an increasing number of diners in Santa Fe have posted on Yelp reports of becoming ill—exhibiting symptoms of food poisoning—after eating at local restaurants. Item 1 requests a summary description of the information represented in a graph displaying the number of such reports during those years. The purpose here is to test our ability to "separate factual information from inferences that might be used to interpret those facts" (Haynes, Lisic, Golts, Stein, and Harris 2016, p. 48). Credit is awarded for stating *just the facts*; *no* credit is awarded when you (also) attribute the increasing number to a possible cause, such as food poisoning; but credit *is* awarded if you (also) claim that the increasing number might be due to a *multiplicity* of possible stated causes. This seems like allowing two wrong answers to make a right: if the goal is to assess students' ability to keep assumed facts *separate* from possible explanatory inferences, then do so consistently.

Item 2 mentions a theory, let's say that the local restaurants are becoming less sanitary and thus more prone to contagions. We are asked to judge to what degree the data represented in the graph support this

theory. And we are told to "explain." Points are awarded for correctly judging how strongly the data support the theory and for suggesting other alternative explanations for the data. By merely being told to "explain," however, we are not fully informed about how robust an answer is expected of us. Stating the number of points possible for this item would be very helpful. This is also true elsewhere in the test, as I will illustrate.

Item 3 asks if there are other "possible explanations" for the data that would not necessarily support the theory (in our case) that local restaurants have been becoming less sanitary and more prone to contagions over the past decade. If you think so, you are to explain by providing a certain number of such alternatives. The problem here is generated by asking for mere *possible* alternative explanations for the data, instead of *plausible* alternatives—namely, explanations that have a *real* possibility, that is a *practically* significant *probability*, of being true and of having enough causal efficacy to produce the data. For example: Yelp is only 14 years old, so the *number* of *earlier* postings about patrons' bad gastric experiences at local restaurants would likely have been fewer and then gradually increase over the decade as the population of Yelp users increases. Or, the city's population or tourism has increased and the *number* of restaurant goers reporting such symptoms could very well have risen without the *proportion* of symptomatic diners doing so.

But among the answers that were deemed acceptable were the likes of these: Perhaps people are increasingly susceptible to intestinal ailments, perhaps people are eating increasingly unsettling foods, or perhaps people are increasingly genetically disposed to exhibiting such symptoms on their own. These explanations are not as far-fetched as that aliens have been increasingly giving local patrons stomach problems (an answer the likes of which was thankfully rejected by my fellow scorers when I asked them at a CAT Training Workshop), but these accepted answers are *mere possible* explanations for which one would *not* have a shred of evidence other than knowing that none of them breaks a law of physics. When I pointed this out at the workshop, I was told that such answers *were* acceptable because "we want to reward creativity." That may be, but creativity should not be rewarded at the expense of *justifiable* empirical reasoning, which is what they are trying to measure in this case.

© Kevin E. Possin.

Item 4 asks what new information would help to evaluate the hypothesis in question. We are asked to find a certain number of such instances and explain how each would help. "Evaluate the hypothesis" is ambiguous: Our evaluations can discuss new information that would help *confirm the hypothesis* (e.g., finding that restaurants in the city increasingly failed annual health inspections over the same decade), or new information that would *disconfirm alternative hypotheses* (e.g., finding that rates of *diagnosed* food poisonings among restaurant goers increased *at medical facilities* in the city during the same decade, while the city's population and tourism had not increased as much), or new information that would help *confirm alternative hypotheses* (e.g., finding that the rate at which people *not* eating at local restaurants reported symptoms of food poisoning had similarly increased over the same decade). The only issue I have with how this item is scored is that a "bonus point" can be awarded for an answer that is explained well; however, no mention of this possibility is made on the test.

Items 5-7 concern the likes of the following scenario: A blogger for a "back to nature" website claims that electromagnetic energy causes prostate cancer. To support his theory, he cities the following evidence: 1) 99% of men diagnosed with prostate cancer live in areas with electrical utilities and appliances, and 2) prostate cancer is rarely reported in areas without access to electricity.

Item 5 requests our judgment about how strongly this evidence supports the blogger's claim—it is disheartening that 20-30% answer this incorrectly (Stein and Haynes 2011, p. 46). Item 6 asks if there are alternative explanations for the evidence besides the blogger's hypothesis, and, if so, to "describe" them. Here again, it would help if we were informed of how many points the answer is worth. And, here again, answers should focus on *plausible common-cause* explanations of the blogger's data. For example, living in an industrialized society with *both* electric utilities and the modern means to: (1) *greater longevity* (so as to more probably eventually contract prostate cancer before dying of other causes), (2) *enhanced medical services* (so as to more probably have one's prostate cancer *diagnosed*), and (3) a wider source of *carcinogens* created in that industrialized culture.

Unfortunately, among the set of accepted answers are the likes of the following: (1) That prostate cancer is partially determined by one's genetic makeup. While this may well explain instances of prostate

cancer, it does *not* explain the blogger's *data*. (2) That the blogger's hypothesis is as spurious as the claim that modern sewage systems are causing prostate cancer. While this argument (*reductio*) from analogy is effective at illustrating how *weak* the blogger's evidence is, it does *not* explain the blogger's *data*. And (3) that there is not a strong causal relationship between electromagnetic energy and prostate cancer, since most men are exposed to electrical utilities and appliances and do not get prostate cancer. While this is true, the question is *not* whether the *relation* is strong—that is, whether the probability of prostate cancer among those exposed to electrical utilities and appliances is high—the question is *not* even whether the blogger's *evidence* is *strong* support for his claim; the question is requesting alternative explanations for the blogger's *evidence,* viz., the high *proportion* [N.B. not correlation!] of exposure to electrical utilities and appliances among men diagnosed with prostate cancer.

Item 7 asks what additional information might help "evaluate" the blogger's claim. Again, we have three options for our answers: We can discuss hypothetical evidence that would *support* or *discredit the blogger's claim*, for example, a controlled or prospective experiment in which an electromagnetic field was administered to one group but withheld from another matched (control) group, to see if there is a statistically significantly higher frequency of prostate cancer in the experimental group. Or we could discuss finding the *statistics to support any of the plausible alternative common-cause explanations* for the blogger's *data* mentioned earlier, to defeat that data's support for the blogger's hypothesis. Or we might discuss finding a physical mechanism or process by which electromagnetic fields cause cancerous mutations in prostate cells. While finding such a mechanism would indeed help *confirm the blogger's hypothesis*, it would be a rather *implausible* discovery in light of the *Earth's* electromagnetic field being so much stronger than those produced by utilities and appliances. So I would rather see someone argue that, if the blogger's hypothesis *were* true, virtually all males would have contracted prostate cancer, given the Earth's comparatively *intense* electromagnetic field (which is 1000 times greater than generated by one's toaster, for example); but they have not, and therefore, the blogger's hypothesis is likely false. Here again, it would also be helpful to know how many possible points our answer is worth. (Fun fact: one of those little refrigerator magnets is 1

million times stronger than that toaster's electromagnetic field—the blogger should be calling for a Surgeon General's Warning!)

Items 8-9 concern the likes of the following scenario: 100 beer drinkers were surveyed while sampling a new India Pale Ale being introduced by the Santa Fe Brewing Company—90% of those surveyed recommended it. Item 8 asks us to judge the degree to which this supports the claim that these beer drinkers think the new IPA is better tasting than the two IPAs already marketed by the Brewery.

Item 9 asks us to provide alternative "interpretations" of what the survey results "could mean," assuming that the survey was properly done and accurate. Up to this point, we have twice been asked for alternative *explanations* of data; but here we are asked for alternative "interpretations" of what the survey data "could mean." 'Interpretation,' however, is more ambiguous than 'explanation.' And what the data could "mean" is likewise ambiguous; for example, it might mean, to someone, that a large proportion of Santa Fe beer drinkers like hoppy beers, which is not the kind of answer the question is testing for. *And*, to ask what the data "could" mean invites providing *merely possible*, but not *plausible*, alternative explanations for the survey data, for example, that most of those surveyed were merely trying to please the Brewery with their answers. *Plausible* alternative explanations could be, for example, that those surveyed thoroughly enjoyed the new IPA but found it no better tasting than the other two IPAs sold by the Santa Fe Brewing Company, probably because they could not distinguish among them; or they recommended it as a great value; or it was a different style IPA, for example with more aromatic hops, making it a refreshing alternative, but not better tasting than the others.

So it should be clear that Part I of the CAT has a few problems:

- The wording of some items is too often a potential source of confusion or distraction.
- We are never explicitly told the maximum points possibly awarded on any item. Usually we can safely presume it, such as when we are told in Item 4 to detail a certain number of cases of additional information that would help evaluate the target hypothesis. But at other times, for example, Item 2, it is not so obvious how robust an answer is expected of us. Perhaps this was intentional—to better test our *disposition* to apply our CT skills. If this were the intention, however, it is

somewhat defeated by prefacing Item 4 with such an explicit reminder that a correlation is difficult to narrow down to a cause and effect relation because a third, *common cause* might be responsible.

- But the most important problem is with how test answers are scored—too often credit is given for answers where credit is not due. (Which makes it all the more depressing that the national average score on the CAT for college seniors is only 19 out of 38 possible—that's a 50%, i.e., a failing grade [Harris et al. 2014, p. 2].)

Part II, on the other hand, is nearly perfect in how it assesses our problem-solving skills in choosing a reason-based action under fixed and contingent conditions. Only Item 14 stands in need of informing us about the maximum points possible, so as to help us determine how robust an answer is expected—with the others, we can reasonably surmise. This portion of the CAT has some similarities with the performance task of the *Collegiate Learning Assessment* [CLA] and now the CLA+ tests. The CAT, however, is *objective* in its acceptable answers, whereas "There is no 'correct' answer [with the CLA+]" (Council for Aid to Education 2017, p. 2)—see (Possin 2013) regarding this problem with the CLA. And the CAT is more robust in its specific assignments (cf. [Council for Aid to Education 2017, pp. 1-15]).

At the risk of appearing obsessive, I want to return to issues involved in scoring the CAT: The CAT scoring process is elaborate, requiring scorers either being trained at a two-day CAT workshop or being trained by someone so trained. There is a scoring manual, with a rubric for each item. Each item is scored by 2 scorers (usually faculty, although the folks at the Center for Assessment and Improvement of Learning will score the CAT for a fee). If those two scorers are in agreement, then that score is assigned; if not, a third scorer decides the majority score or the average score (if none agree).

While agreement between the first and second grader in faculty scoring sessions is high (R = .92), this reliability does not ensure "validity." Upon rescoring samples (15-20%) of tests, it was found that the average "error rate" (i.e., noncompliance with the answer key) in scoring sessions is 5.4% (N = 280 sessions, 14,600 tests) (Stein et al. 2016, p. 5). That is 2 points out of the 38 possible, which seems more significant in light of the national average score for seniors being 19

and the mean 4-year gain being 3.9 points (N = 33,000) (Harris et al. 2014, p. 2). This, and the frequent variability in the members of scoring sessions, explains why we are advised to use the CAT to assess gains regarding *groups* rather than *individuals* (Center for Assessment and Improvement of Learning 2019, p. 24).

Scoring the CAT is also very labor intensive, with each test item being scored by two faculty and then a third if there is a disagreement. According to the training manual (ibid, p. 25), novice scorers should score 6-7 tests per day-long session, experienced graders 10-14, and very experienced graders about 20 tests per session. This means that you often cannot have every student in a course or a program pre-/post-tested and scored using the CAT to determine *individual* gains in critical-thinking skills. The recommended workaround is to have all subjects tested but score only a random sample of at least 10-15, but preferably about 30, and statistically determine individual gains for those students to make a judgment about *group* gains. Machine scoring of the CAT might well resolve this limitation and this problem with reliable scoring: A sample of 500 tests was used to compare machine scoring to "expert" scoring; the percentage error of the former was 1.37%, well within the accepted 5% limit set for institutional scoring sessions (Center for Assessment and Improvement of Learning 2018).

While machine scoring might solve these two problems with using the CAT, it would cancel a much-heralded virtue of having faculty not only administer but also score the CAT: They get to see first-hand where their students are lacking in critical-thinking skills. Faculty can then use the CAT results not only as a means of accountability but also as a means to adjusting their curriculum so as to improve the critical-thinking skills of their students, in turn verifying this improvement by means of in-class practice exercises, course exams, and critical-thinking assessment post-course and in future courses.

3. CAT Apps

The CAT is to be used as a "model" for creating these course exercises and exams; faculty are to use the CAT's items as templates, clothing them in the "content and methods" of their specific courses; faculty are to develop their own discipline-specific "analogous activities" (Stein and Haynes 2011, p. 48). Half of the CAT Training Workshop is now

dedicated to this topic of creating these "CAT Apps." For what it's worth, 45% of the faculty attending these CAT Apps Workshops *self-reported* changing their curriculum in this way (or *intending* to), albeit no testing was done to verify student gains resulting from such curricular changes (Haynes et al. 2016, pp. 52-6). Faculty find it especially difficult to develop reasonable scoring rubrics for their own CAT Apps (Haynes, Lisic, Harris, Leming, Shanks, and Stein 2015, p. 43). Which is unfortunate, because, as we have seen, valid scoring is *a necessary condition* for meaningful assessment. Indeed, if the scoring of these CAT Apps exercises and exams is *incorrect*, faculty could be doing more harm than good towards enhancing their students' critical-thinking skills.

At those institutions cited as introducing new courses or curricular changes that emphasize critical thinking and produce "significant gains" on the CAT, it is *not* always clear that the course curricula were made up significantly of CAT Apps that could explain those gains (Stein and Haynes 2011, pp. 48-9). For example, the general education Foundations of Science course created at Sam Houston State University—dedicated to the study of generic critical-thinking skills involved in scientific reasoning and reaping average effect-size gains of 0.73 on the CAT—used two texts and numerous case studies, but no CAT Apps were mentioned (Rowe, Gillespie, Harris, Koether, Shannon, and Rose 2015). A similar analysis applies to (Gottesman et al. 2013), another example cited by Haynes et al. (2015). However, a genuine instance of using CAT Apps after attending a CAT Apps workshop is discussed in (Carson 2015) (Haynes et al. 2015, p. 44). The scenarios, case studies, and exercises Carson designed for her first-year biology course were very admirable, and her 14 students experienced a statistically significant effect-size gain of 0.67 on the CAT. (For another instance in which the use of CAT Apps seems to have had an impact on post-CAT gains, see [Styers, Van Zandt, and Hayden, 2018].)

But let's take a step back and look at this recommendation to use CAT Apps: We are to draft "analogous" versions of the CAT's scenarios and questions, *using different, discipline-specific, contents*, thus providing in-class practice exercises and exams for the students before they take the post-course CAT. CAT Apps are "models" and "adaptations" "derived from" the CAT; faculty are to "adapt questions" from the CAT (Haynes et al. 2015). Prompts for the CAT Apps (as

suggested in the CAT Apps worksheets used at the CAT Training Workshops) are *virtually identical to the prompts used in the CAT itself*. (See, e.g., [Schmidl-Gagne, Lisic, and Harris 2018].) This seems to be "teaching to the test." If indeed students show gains on the CAT after such activities, it should come as no surprise. Stein, Haynes, and Redding (2007, pp. 295-6) discuss how students trained for six hours on how to *score the CAT itself, using the official scoring manual*, did significantly better on the CAT than a control group did. That is hardly surprising. They also did better on a set of analogous "transfer" questions. Again, not a surprise. And ditto for the *reverse*, one should expect.

Haynes et al. (2015) cite (Gibbs and Simpson 2004) in their defense, claiming that "assessment guides the students' prioritization of what is important for success in a particular class" (p. 41). This is just a fancy way of saying that students only study and learn what they think will be on the test or, more generally, what will determine their grade. There is no need to cite research to support this fact—years of students' asking "Is this going to be on the test?" is evidence enough. But CAT Apps turn this regrettable "is" into an "ought." Haynes et al. recommend, in essence, that faculty have the tail wag the dog, or rather the CAT wag the dog.

Luckily, the CAT *test per se* has enough content validity that teaching *to* it via CAT Apps could indeed enhance its targeted sliver of critical-thinking skills in students. It's just that when students slated to take the post-course CAT have practiced its items using doppelgangers, it is not great cause for congratulations all around when those students show modest gains in their scores.

Post-test gains on the CAT are also claimed to be evidence for the "concurrent validity" of the CAT Apps *and vice versa* (Center for Assessment and Improvement of Learning 2019, p. 32). This circular validation, however, is an *abuse* of that species of criterion validity—one of the indicators in this correlative relation must get its validity *elsewhere*.

Another problem with focusing on CAT Apps that only "adapt questions" from the CAT, is that students never see an instance in which *the evidence in fact strongly supports an hypothesis by virtue of the latter being the best explanation of the former*. By discussing only cases in which there are possible/plausible alternatives to causal and

statistical hypotheses, one invites students to become skeptical of evidence-based claims in general, especially when students are erroneously encouraged to be "creative" and think that the *mere possibility* of an hypothesis' being false (given the evidence) is grounds for rejecting that hypothesis.

And finally, what about the care and feeding of the CAT? Its annual fee for use is $300. Attending a CAT Training Workshop is $550, plus hotel and transportation. Each CAT test booklet is $9.95, and each online test is $20. And faculty scorers need to be compensated or the Center for Assessment and Improvement of Learning needs to be paid for doing the scoring. So, pre/post-testing for a class of 30 would *start* at $1500 or so, assuming the scorers are working for nothing, making the CAT rather unaffordable for most faculty and departments.

4. Conclusion

So, would I use the CAT in my critical thinking course? I would, but only if that course was dedicated to the study of scientific and elementary-statistical reasoning and problem solving, and only if my departmental budget was quite substantial, and only if I could change the CAT and the scoring manual in the ways I have discussed.

Acknowledgements

Portions of this article appear in "Thinking Critically about Critical Thinking Assessment" coauthored with Donald Hatcher, in *Critical Thinking and Reasoning: Theory, Development, and Practice*," (2020) edited by Daniel Fasko and Frank Fair.

References

Alvarez, C., Taylor, K., and N. Rauseo. 2015. Creating thoughtful salespeople: Experiential learning to improve critical thinking skills in traditional and online sales education. *Marketing Education Review* 25(3): 233-243.

Carson, S. 2015. Targeting critical thinking skills in a first-year undergraduate research course. *Journal of Microbiology and Biology Education* 16(2): 148-156.

Center for Assessment and Improvement of Learning. 2015. *CAT training manual.*

Center for Assessment and Improvement of Learning. 2016. CAT instrument technical information. URL accessed 8 April 2020: https://www.tntech.edu/cat/pdf/reports/CAT_Technical_Information_V8.pdf

Center for Assessment and Improvement of Learning. 2018. Machine scoring of student responses on the CAT. URL accessed 8 April 2020: https://www.tntech.edu/cat/reports.php

Center for Assessment and Improvement of Learning. 2019. *CAT training manual.*

Council for Aid to Education. 2017. CLA+ sample assessment. URL accessed 8 April 2020: https://cae.org/images/uploads/pdf/CLA_Practice_Assessment.pdf

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

Gasper, B., Minchella, D., Weaver, G., Csonka, L., and S. Gardner. 2012. Adapting to osmotic stress and the process of science. *Science* 335(6076): 1590-1.

Gibbs, G. and C. Simpson. 2004. Conditions under which assessment supports students' learning. *Learning and Technology in Higher Education* 1: 3-30.

Gottesman, A. and S. Hoskins. 2013. CREATE Cornerstone: Introduction to scientific thinking, a new course for STEM-interested freshman, demystifies scientific thinking through analysis of scientific literature. *CBE-Life Sciences Education* 12(1): 59-72.

Harris, K., Stein, B., Haynes, A., Lisic, E., and K. Leming. 2014. Identifying courses that improve students' critical thinking skills using the CAT instrument: A case study. *Proceedings of the Tenth Annual International Joint Conference on Computer Information, Systems Science, and Engineering*.

Hatcher, D and K. Possin. 2020. Critically thinking about critical thinking assessment. In *Critical thinking and reasoning: Theory, development, and practice,* eds. D. Fasko and F. Fair. The Netherlands: Brill-Sense.

Haynes, A., Lisic, E., Goltz, M., Stein, B., and K. Harris. 2016. Moving beyond assessment to improving students' critical thinking skills: A model for implementing change. *Journal of the Scholarship of Teaching and Learning* 16(4): 44-61.

Haynes, A., Lisic, E., Harris, K., Leming, K., Shanks, K, and B. Stein. 2015. Using the *Critical Thinking Assessment Test* (CAT) as a model for designing within-course assessments: Changing how faculty assess student learning. *INQUIRY: Critical Thinking Across the Disciplines* 30(3): 38-48.

Kahneman, D. 2011. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kruger, J. and D. Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77(6): 1121-1134.

Lawshe, C. 1975. A quantitative approach to content validity. *Personnel Psychology* 28: 563-575

Paul, R., Elder, L., and T. Bartell. 1997. Study of 38 public universities and 28 private universities to determine faculty emphasis on critical thinking in instruction. URL accessed 8 April 2020: https://www.criticalthinking.org/pages/study-of-38-public-universities-and-28-private-universities-to-determine-faculty-emphasis-on-critical-thinking-in-instruction/598

Possin, K. 2002. *Critical thinking*. Winona, MN: The Critical Thinking Lab.

Possin, K. 2008. A field guide to critical-thinking assessment. *Teaching Philosophy* 31(3): 201-28.

Possin, K. 2013. Some problems with the *Halpern Critical Thinking Assessment* (HCTA) *Test*." *INQUIRY: Critical Thinking Across the Disciplines* 28(3): 4-12.

Possin, K. 2013. A serious flaw in the *Collegiate Learning Assessment* [CLA] *Test*." *Informal Logic* 33(3): 390-405.

Possin, K. 2014. Critique of the *Watson-Glaser Critical Thinking Appraisal Test*: The more you know, the lower your score. *Informal Logic* 34(4): 393-416.

Rowe, M., Gillespie, M., Harris, K., Koether, S., Shannon, L., and L. Rose. 2015. Redesigning a general education science course to

promote critical thinking. *CBE-Life Science Education* 14(3): 1-12.

Schmidl-Gagne, K., Lisic, E., and K. Harris. 2018. Integrating critical thinking to produce successful student civic engagement using the CAT framework. *American Association of Colleges and University Conference, General Education and Assessment: Foundations for Democracy*. URL accessed 8 April 2020: https://www.tntech.edu/cat/pdf/presentations/2018_AACU_workshop.pdf

Stein, B. and A. Haynes. 2011. Engaging faculty in the assessment and improvement of students' critical thinking using the Critical Thinking Assessment Test. *Change: The Magazine of Higher Learning* March: 44-49.

Stein, B., Haynes, A., and M. Redding. 2006. Project CAT: Assessing critical thinking skills. *Proceedings of the National STEM Assessment Conference-NSF and Drury University*.

Stein, B., Haynes, A., and M. Redding. 2016. National dissemination of the CAT instrument: Lessons learned and implications. *Proceedings of the AAAS/NSF Envisioning the Future of Undergraduate STEM Education: Research and Practice Symposium*. URL accessed 8 April 2020: https://www.tntech.edu/cat/pdf/reports/AAAS_2016_Stein_Haynes_Redding.pdf

Stein, B., Haynes, A., Redding, M., Ennis, T., and M. Cecil. 2007. Assessing critical thinking in STEM and beyond. In *Innovations in e-learning, instruction technology, assessment, and engineering education*, ed., M. Iskander, 79-82.

Stein, B., Haynes, A., Redding, M., Harris, K., Tylka, M., and E. Lisic. 2010. Faculty driven assessment of critical thinking: National dissemination of the CAT instrument. *Proceedings of the 2009 International Joint Conference on Computer, Information, and Systems Sciences, and Engineering*.

Styers, M., Van Zandt, P., and K. Hayden. 2018. Active learning in flipped life science courses promotes development of critical thinking skills. *CBE-Life Science Education* 17(3): 1-13.

Wittgenstein, L. 1953. *Philosophical investigations*. New York, NY: Macmillan Company.