Critique of the *Watson-Glaser Critical Thinking Appraisal Test*: The More You Know, the Lower Your Score

KEVIN POSSIN

Professor Emeritus, Philosophy
Winona State University
The Critical Thinking Lab
24847 County Road 17
Winona, MN 55987
USA
kpossin@winona.edu

Abstract: The *Watson-Glaser Critical Thinking Appraisal Test* is one of the oldest, most frequently used, multiple-choice critical-thinking tests on the market in business, government, and legal settings for purposes of hiring and promotion. I demonstrate, however, that the test has serious construct-validity issues, stemming primarily from its ambiguous, unclear, misleading, and sometimes mysterious instructions. Erroneously scored items further diminish the test's validity. As a result, having enhanced knowledge of formal and informal logic could well result in test subjects receiving lower scores on the test. Many of the W-G's validity issues, however, could be easily remedied.

Résumé: Le *Watson-Glaser Critical Thinking Appraisal Test* est un des tests de la pensée critique à choix multiples les plus anciens et les plus fréquemment utilisés à des fins de recrutement et de promotion par des entreprises, le gouvernement et divers domaines juridiques. Je démontre, cependant, que le test a des problèmes graves provenant principalement de ses instructions ambiguës, confuses, trompeuses, et parfois mystérieuses. Le fait que des questions sont mal notées diminue davantage la validité du test. En conséquence, des sujets doués d'une bonne connaissance des logiques formelle et non formelle qui subissent l'épreuve pourraient bien aboutir à des résultats plus faibles que les résultats des sujets moins doués de telles connaissances. Plusieurs questions de validité du test W-G, cependant, pourrait être facilement résolues.

Keywords: arguments, construct validity, content validity, critical-thinking assessment, critical-thinking skills, deductive and inductive reasoning, syllogistic reasoning, *Watson-Glaser Critical Thinking Appraisal Test*

1. Introduction

The *Watson-Glaser Critical Thinking Appraisal Test* [W-G] is one of the oldest multiple-choice critical-thinking [CT] tests on

the market. Despite the introduction of many competing means of CT assessment, many focusing on "performance" or "constructed response" tasks as opposed to "recognition" tasks, the W-G has not only survived but evolved and thrived. It has long been (Ryan & Sackett, 1987), and continues to be, the most frequently used CT assessment test internationally, in business, government, and legal settings, for purposes of hiring and promotion. A version of the W-G—the *Bar Course Aptitude Test* (2013)—is now becoming mandatory in the UK, to determine admission into law schools. So, just how good is the W-G at assessing one's CT skills? That is the important question here. Please note that this will not be a traditional psychometric critique of the W-G. As it turns out, problems with the validity of the W-G become most apparent when viewed from a different perspective—that of the subject matter of critical thinking itself.

2. Test Format

The W-G was first developed in the 1930s and has undergone numerous revisions. Two early versions of it, Forms A and B (1980), are still in use today, have 80 items, and take approximately 55 minutes to complete. The W-G Short Form (1994) consists entirely of a subset of Form A, has 40 items, and takes approximately 35 minutes to complete. The more recent Forms D and E (2009) partially consist of subsets of Forms A and B respectively, have 40 items, and take approximately 35 minutes to complete. These three shorter versions are available in the U.S. online. Their items have been updated to discuss more contemporary topics, with Form D focusing most on business topics. The most recent online version, the W-G Unsupervised (2011), is uniquely assembled from pools of items for each test subject, so as to make the test more secure while (supposedly) still keeping all generated tests comparable in difficulty.

Pricing for access to the W-G varies: $25-35 per test, depending on the type of test and its format. A scoring "Profile Report" is available for approximately $30. It includes the test subject's raw score and overall performance percentile; subscale scores and performance percentiles in three categories of CT skills (recognizing assumptions, evaluating arguments, and drawing conclusions); comparative ratings to other individuals in the test subject's norm group; and a brief description as to

what the subject's scores suggest with respect to CT-skill levels. (Information on prices and availability is scattered and subject to change, which is why I forego citations and merely refer the reader to the websites of Pearson Assessment, the publisher of the W-G, and TalentLens, its distributer.)

3. Test Content

When attempting to craft a test for measuring CT skills, everything depends first on what one thinks constitutes CT. People's definitions of CT have become so elastic that almost anything passes for it, making the notion of CT nearly worthless. For a discussion of this unfortunate state of affairs, see, e.g., (Possin, 2008, pp. 203-06; Ennis, 2013, pp. 30-34). Fortunately, Watson and Glaser kept a tighter rein on their concept of CT, which

> involves three things: (1) an attitude of being disposed to consider in a thoughtful way the problems and subjects that come within the range of one's experiences, (2) knowledge of the methods of logical inquiry and reasoning, and (3) some skill in applying those methods. Critical thinking calls for a persistent effort to examine any belief or supposed form of knowledge in the light of the evidence that supports it and the further conclusions to which it tends. It also generally requires the ability to recognize problems, to find workable means for meeting those problems, to gather and marshal pertinent information, to recognize unstated assumptions and values, to comprehend and use language with accuracy, clarity, and discrimination, to interpret data, to appraise evidence and evaluate arguments, to recognize the existence (or non-existence) of logical relationships between propositions, to draw warranted conclusions and generalizations, to put to test the conclusions and generalizations at which one arrives, to reconstruct one's patterns of belief on the basis of wider experience, and to render accurate judgments about specific things and qualities in everyday life. (Glaser, 1941, pp. 5-6)

In light of this admirable analysis of CT, all versions of the W-G evenly address the following categories of CT skills:

1. *Making inferences*: Correctly judging whether a conclusion is "definitely true," "probably true," "probably false," or "definitely false," based on a set of premises, or whether there are "insufficient data" to draw a conclusion.
2. *Recognizing assumptions*: Correctly judging whether an assumption is necessary or not for the truth of a statement.
3. *Reasoning deductively*: Correctly judging whether a conclusion is a logical implication of a set of premises.
4. *Interpreting arguments*: Correctly judging whether a conclusion "follows beyond a reasonable doubt" from another statement.
5. *Evaluating arguments*: Correctly judging whether an argument is "strong" or "weak."

The W-G, therefore, correctly focuses on some of the most crucial CT skills used in the assessment of both deductive and inductive reasoning, and it tests for almost all of the CT skills that, for instance, Menkes (2005) argues should be the focus of any executive hiring search.

4. Validity

The validity of a test is essentially a matter of how well the scores on the test are an accurate measurement of what the test is designed to measure. There are various aspects of validity, however; the primary of which is *content* validity. This is how well the test acts as a gauge that accurately measures, in this case, *real* CT skills. Just as one's fuel gauge was designed to measure the level of fuel in one's tank, so the W-G was designed to measure test subjects' level of CT competency.

A gauge, however, might accurately indicate that one's tank is empty and yet the person using the gauge might misinterpret its reading, e.g., thinking that the needle's being on 'E' means that the tank contains "enough." This brings us to the concept of *construct* validity, whereby the subjects' *assigned scores* on their test answers must accurately represent the subjects' implementation of real CT skills. If the *accepted* answers used to score a content-valid test are erroneous, construct validity diminishes.

An accurate gas gauge which is properly read not only informs you as to the level of fuel in the tank but also enables you to make other inferences or predications, e.g., that you will need to stop at the next exit to refuel. The fact that it took quite a large quantity to fill your tank when the gauge's needle was on 'E' is *modest* evidence of the gauge's accuracy (*modest*, because the tank could be empty while the gauge is broken, *stuck* on 'E'). And if you miss that next exit and, to your surprise, are able to drive on for many more miles, you would have rather good evidence that your gauge is not accurate after all. This illustrates the notion of a test's *criterion* validity. Finding a correlation between subjects' test scores and another predicted variable, supposedly correlated to what the test measures, is modest evidence of the test's (criterion) validity.

5. W-G Validity

As I mentioned earlier, I think the W-G properly focuses on some of the crucial categories of CT skills. But is its degree of content validity just a happy accident? What is the *evidence* for the test's content validity, confirming that its items indeed test for those CT skills? According to the *W-G User-Guide and Technical Manual*:

> W-GCTA passages contain stimulus material similar to that encountered on a daily basis at work, in the classroom, and in newspaper or magazine articles. Respondents are required to show critical thinking in identifying valid and invalid inferences from passages, identifying underlying assumptions and evaluating the strength of arguments. Therefore the nature of the task is that it will require critical thinking with relevant contextual material. (2012, p. 38)

This just begs the question, however, to say that it is simply in "the nature of the task" (of taking the test) that one is required to use real CT skills. A standard approach to verifying content validity, developed by Lawshe (1975), is basically to get consensus approval of the test's content and answers *from an army of subject-matter experts*. I find no mention of such a study anywhere; so let me enlist in that army right now and begin that study. The *Manual* (2012, p. 26) states that items prepared for the newest version of the W-G "were written by

experienced item writers, occupational psychologists or psychometricians with at least 15 years experience." However, these are *not* the credentials of *subject-matter experts in CT*.

To argue for the construct validity of the W-G, the *Manual* (2012, pp. 24-26) appeals to factor analysis studies regarding correlations among the sections of the W-G. While they might confirm the internal structure of the test's categorization of items (e.g., indicating the appropriateness of combining the sections on Inference, Deduction, and Interpretation into the new subscale-scoring category of Drawing Conclusions), they indicate at most a degree of *reliability*, which is necessary, but not sufficient, for the test's validity.

The evidence provided in the *Manual* (2012, pp. 41-42) and in (Ejiogu, Yang, Trent, & Rose, 2012) for the W-G's criterion validity makes it understandable why the test is being used so widely for hiring and promotion purposes. But demonstrable correlations between W-G scores and job success may very well hold because of a correlation between job performance and, e.g., mere test-taking abilities in general. Evidence for criterion validity is notoriously poor at screening out alternative causal explanations. As a result, it provides at most an installment in arguing for overall validity.

If that is the evidence offered *for* the W-G's validity, what is the evidence offered *against* it?

John McPeck (1981;1984) provides the most global critique of the W-G possible: The W-G fails to validly assess test subjects' general CT skills because there simply are no such generic CT skills to assess.

> [CT] is the appropriate use of *reflective scepticism* within the problem area under consideration…. Since critical thinking is always 'critical thinking about X', it follows that critical thinking is intimately connected with other fields of knowledge. Thus the criteria for the judicious use of scepticism are supplied by the norms and standards of the field under consideration. (1981, pp. 7-8)

McPeck's errors here are glaring, however: If his account of CT were correct, one could not use CT to reason to any *positive* conclusion about what one *ought* to believe. And his argument would imply, e.g., that I've never learned to tie shoes, since I've only learned to tie leather shoes, tennis shoes, brown shoes, etc.,

with cotton laces, leather laces, nylon laces, etc. But that's absurd. Moreover, for CT skills to be domain specific, there must be specific domains, each with their own CT skills, none of which are transferable. But that too is absurd—the epistemological universe does not come in such tidy and independent categories. For example, learning the basic considerations of using experimental evidence to argue for or against an hypothesis in one area of science helps one to use empirical evidence to do likewise in other disciplines too, especially in other sciences. And denying the consequent is certainly a relevant and valid argument form no matter what its subject matter; so being able to recognize it as such is a very useful and transferable CT skill.

A general problem regarding very early versions of the W-G was raised by Robert Ennis (1958), who pointed out that a "pathological doubter," viz., a skeptic about whether anything can truly be said to function as a good reason for believing anything else, would do quite well by answering merely from their grim ideology. Fortunately, this criticism no longer applies to currently used forms of the W-G; for example, with Forms A and B, such a skeptic would score only .45 and .46, respectively.

While the W-G identifies and tests for *some* of the crucial categories of CT skills, it has some crucial omissions. For instance, the W-G fails to include any assessment of the ability to identify informal fallacies, i.e., the classic rhetorical tricks of argumentation. Please note that I am *not* suggesting that test subjects should be expected to identify informal fallacies *by name*, only that subjects should be tested on their ability to recognize that an argument in question is fallacious; e.g., they should correctly identify when someone is illegitimately using a word in two different ways, but they need not know that this is called 'equivocation.' Among the test's other notable omissions, first spotted by Govier (1987, p. 256), are arguments by analogy, formal fallacies, and the use (or abuse) of definitions. Because of this, the W-G has a content validity issue—it inadequately detects some crucial CT skills, because it makes no *attempt* to detect them. While this is to be expected to some extent, since no CT assessment test can address *all* CT skills, shortening the test to 40 items only makes matters worse in this respect.

It should be noted that the W-G has long been characterized (and criticized) as being focused on only deductive reasoning while omitting inductive reasoning—e.g.,

by Govier (1987, p. 256) and the Hunt Executive Search service (2013). But this is a serious misrepresentation, as I will soon demonstrate.

The overall format of the W-G is also problematic in that only one of the five sections consists of test items with five-option answers; the rest of the test consists of items with only two-option answers. This dramatically increases the probability of lucky guesses, thereby decreasing the test's sensitivity and its capacity to measure CT skills and their improvement by means of pre- and post-testing.

There are numerous issues that arise more specifically within each of the five sections of the W-G. I will discuss them in order.

5.1 Inference

In the directions for this section, the test subject is told the following:

> An inference is a conclusion a person can draw from certain observed or supposed facts….
>
> *In this test, each exercise begins with a statement of facts that you are to regard as true.* After each statement of facts, you will find several possible inferences—that is, conclusions that some person might draw from the stated facts…. [Answer]
>
> T if you think the inference is definitely TRUE; that it properly follows beyond a reasonable doubt from the statement of facts given.
>
> PT if, in the light of the facts given, you think the inference is PROBABLY TRUE; that it is more likely to be true than false.
>
> ID if you decide that there are INSUFFICIENT DATA; that you cannot tell from the facts given whether the inference is likely to be true or false; if the facts provide no basis for judging one way or the other.
>
> PF if, in the light of the facts given, you think the inference is PROBABLY FALSE; that it is more likely to be false than true.

> F if you think the inference is definitely FALSE; that it is wrong, either because it misinterprets the facts given, or because it contradicts the facts or necessary inferences from those facts.
>
> Sometimes, in deciding whether an inference is probably true or probably false, you will have to use certain commonly accepted knowledge or information that practically every person has.

These directions are by no means "clear enough" (Fisher and Scriven, 1997, p. 117); they are quite confusing for numerous reasons. And the extent to which they are in turn misleading test subjects from answering correctly, they are diminishing the test's construct validity. First, an inference is *not* a conclusion; it is a mental process wherein one believes a conclusion on the basis of one's belief of the premises. Hence, *inferences* are neither true nor false; they are valid or invalid, justified or unjustified, but not true or false. *Statements* are what can be true or false. So, the instructions should be directing one to mark "T" if one thinks the *conclusion* is true, etc. McPeck (1981, pp. 134-5) makes much the same criticism.

Another confusion is initiated with the use of the word 'definitely.' Normally, if one says that a statement is *definitely* true, one is claiming that it is *certainly* true—its probable truth is 1. (For example, if I buy a lottery ticket, I am *not definitely* going to lose; whereas, if I don't buy any ticket, I am definitely not going to win.) To clarify what is meant by a conclusion's being "definitely" true, the directions state that the conclusion "follows beyond a reasonable doubt." Those familiar with formal logic use the phrase 'follows from' to describe the relation of *entailment* between the premises and conclusions of *deductive* arguments—i.e., arguments in which, given the truth of the premises, the conclusion is *necessarily* true. But this understanding of the directions is in conflict with the phrase 'beyond a reasonable doubt.' *That* phrase is standardly used to describe *inductive* arguments (e.g., in legal settings) in which a conclusion is *very probably true* (but *not necessarily true*), based on one's premises. All this leaves test subjects reasonably wondering if they should be judging the exercises by the standards of *deductive* reasoning or the standards of *inductive* reasoning, *when, in fact, they should be doing the latter*. This

could be especially confusing to someone who has taken a CT or introductory logic course, in which they have studied the important differences between inductive and deductive arguments and their respective cogency conditions. If someone, as a result, mistakenly concludes that this section of the test is dedicated to the assessment of *deductive* inferences, then having enhanced CT skills will reduce their score.

It is ironic (but understandable) that this error of applying deductive standards of reasoning to the test's inductive arguments is committed even more by *critics* of the W-G than by the test's authors. For example, Norris and Ennis (1989, pp. 58-9) and Fisher and Scriven (1997, pp. 207-8) discuss the scenario on Form A about Mr. Brown, who was brought before a municipal court for the sixth time in a month on a charge of keeping his pool hall open after 1 a.m. He admitted his guilt and was fined $500, as in each prior instance. The test scores it as PT that it is sometimes to Brown's advantage to keep his pool hall open after hours even at risk of the fine. The four critics object, claiming that the answer should be ID, because they can imagine so many possible alternative explanations for why Brown is keeping his pool hall open illegally. The answer, they insist, depends on how imaginative the test subject is. But that can't be right: Just because one can *imagine* a *possible* alternative explanation, based on one's personal experience, doesn't mean that it is a *plausible* alternative explanation. All the far-flung possible explanations for Brown's behavior that they suggest (e.g., that it was the Mafia or Brown's son or wife, not Brown, who was to blame) don't add up to a *real* possibility. And none of their alternative explanations explains why Brown so readily admits his guilt.

This last little fact, stated in the scenario, is also telling against one of the test's own answers: It scores as ID the statement that Brown repeatedly flouted the closing law in hopes of getting it repealed. But that can't be right: The degree to which it was probably true that Brown is finding it advantageous to stay open after hours (despite its added cost of doing business) is the degree to which it is probably *false* that his motive is one of protesting the law. Moreover, he isn't trying to get publicity by, e.g., getting charged on *successive* nights in the presence of the news media. And, by readily admitting his guilt and paying his fines, he isn't using the Court as a soapbox for objecting to the law.

Fawkes et al. (2001, pp. 22-25) are the latest to join the chorus making this hypercriticism. They argue that many items

in this section should be scored as ID by repeatedly manufacturing alternative *logically* possible assumptions and conclusions that are (admittedly) "not inconceivable" (p. 23) but which are simply not plausible and, therefore, not relevant to undermining inductive arguments.

A third issue arises because the test subject is given ambiguous directions about what the inferences to the conclusions in question are to be based on—*solely* on "the statement of facts given" or also on "commonly accepted knowledge"? This ambiguity could have been avoided by stating in the initial paragraph of the directions that the test subject will sometimes be required to assess the probable truth or falsity of conclusions on the basis of common background knowledge.

But two problems would still remain: The background beliefs of some test subjects will reasonably differ and result in different, but equally justifiable, judgments about the probability of the conclusions drawn on those differing bases (Fisher and Scriven concur, [1997, p. 117]). And the degree to which one is judging the probability of a conclusion on the basis of one's own background beliefs is the degree to which one is *not* judging its probability on the basis of the statements provided in the scenarios—which is the expressed project in this section of the test, after all. These issues are illustrated in the example provided in the directions:

> Two hundred students in their early teens voluntarily attended a recent weekend student conference in a Midwestern city. At this conference, the topics of race relations and means of achieving lasting world peace were discussed, since these were the problems the students selected as being most vital in today's world.
>
> 1. As a group, the students who attended this conference showed a keener interest in broad social problems than do most other students in their early teens.
> 2. The majority of students had not previously discussed the conference topics in their schools.
> 3. The students came from all sections of the country.
> 4. The students discussed mainly labor relations problems.

5. Some teenage students felt it worthwhile to discuss problems of race relations and ways of achieving world peace.

#1 is scored as PT, based *solely* on the assumption that early teens are generally apathetic about such broad social problems. It is not scored as T, because "it is possible that some of the students volunteered to attend mainly because they wanted a weekend outing." This mere possibility, however, is perfectly compatible with #1 being true beyond a reasonable doubt; so the answer *should* have been T. This indicates that the authors of the W-G are themselves guilty of misapplying the standards of deductive reasoning to the inductive argument involving #1. #2 is scored as PF, *solely* on the basis of the background belief that the teenagers' awareness of the topics probably stemmed from discussions with teachers and classmates. #3 is scored as ID, for lack of evidence. But I would argue that the answer *should* be PF: Early teens don't readily have the resources and privileges to fly in from the Coasts to attend a Midwestern conference. #4 is scored as F, because of the stated facts in the scenario. #5 is scored as T, because it "necessarily follows from the given facts." This answer is correct, but this *reason* is incorrect: It is still *logically possible* that #5 is *false* given the facts, even though #5 is true beyond a reasonable doubt (given those facts and the reasonable assumption that people deem the topics they select worthwhile). So, the answer to one of these five items is affected by a difference in background beliefs, and even the authors of the W-G (twice!) erroneously apply the standards of deductive reasoning to the inductive arguments they are scoring on their test.

A test item in which one's background belief could very well affect one's answer occurs on Form A, concerning a scenario in which, after all the labor unions joined a community's Chamber of Commerce, the unions' members "worked with representatives of other groups on committees, spoke their minds, participated actively in the civic improvement projects, and helped the Chamber reach the goals set in connection with those projects." On the basis of this, it is scored as PF that "Most of the union representatives regretted having accepted the invitation to participate in the Chamber of Commerce." This seems correct. But, on the same basis, it is scored as ID that "Some of the Chamber of Commerce members came to feel that their president had been unwise in asking the

union representatives to join the Chamber." This seems incorrect: Anyone who is the least bit familiar with the usual ideological differences between business owners and labor unions would reasonably believe that it's very likely that *at least one* Chamber member is still not keen about the extensive union membership in the Chamber.

McPeck (1981, pp. 135-7) makes a similar criticism, suggesting that items in this section should have been made "self-contained," with test subjects instructed to judge the inferences *only* on the basis of the statement of facts given. However, McPeck's criticism devolves into a strawman when he claims that it would be reasonable for test subjects who have studied statistics to become the likes of Ennis' "pathological doubters," scoring all of the items involving inductive inferences as ID, for lack of specific data sets. Fawkes et al. (2001, p. 22) similarly overreact, claiming that "none of the results of the test can be trusted."

In summary: While this section of the W-G is not as flawed as some critics have claimed, its directions are causing well-published authors in CT to misconstrue the task, indicating that the directions are in need of revision and the construct validity of the section is suspect.

5.2 Recognition of assumptions

In this section, one's task is to judge whether a proposed assumption was *necessarily* made or not. However, the directions provided do not make this clear:

> Below are a number of statements. Each statement is followed by several proposed assumptions. You are to decide for each assumption whether a person, in making the given statement, is really making that assumption—that is, taking it for granted, justifiably or not.
>
> If you think that the given assumption is taken for granted in the statement, [answer] "ASSUMPTION MADE"…. If you think the assumption is *not* necessarily taken for granted in the statement, [answer] "ASSUMPTION NOT MADE."

The ambiguity here is in determining whether an assumption is *made*—is the assumption being *in all probability*

made or is it being *necessarily* made? The test directions explicitly disambiguate this issue for whether the assumption is *not* made—it is when the assumption is not *necessarily* taken for granted. But it is only in the second item of the test's example that the instructions are disambiguated for determining when the proposed assumption *is* made by the target statement: The example's answer is explained by the remark: "(This is necessarily assumed in the statement since, in order to save time by plane, it must be possible to go by plane.)" This ambiguity should not have occurred in the first place, and its correction should not have been buried in a parenthetical remark that test subjects can easily fail to appreciate.

This section of the test, then, is dedicated to assessing one's ability to recognize *deductive* relations among statements. Having settled this, I must challenge the accepted answer to one of the items in this section of Form A: Someone states that they are going to South America and want to be sure they don't get typhoid fever, so they decide to go to their physician for a vaccination before the trip. Do they thereby assume that typhoid fever is more common in South America than it is where they live? This would *seem* to be so, especially in light of their expressed desire to be *certain* they don't get typhoid fever. And that is the way it is scored. But it is still *logically* possible that typhoid fever is as common where they live and that they are well aware of this—they are just especially interested in avoiding contracting it while in South America because, let's say, they don't want to get sick on their very expensive trip. This latter answer might be characterized as nitpicking rather than recognizing assumptions, but it is still a *logical possibility* and, therefore, the assumption is *not logically necessary*, albeit it *is* enormously *probable*. This item is reused on the Short Form, but not on Form D.

A more controversial case arises on Form B, in which someone says, "If you don't believe me, I'll prove it to you logically." The speaker is scored as necessarily assuming that presenting a logical proof will alter, and influence a change in, your belief on the matter. I would be the first to admit that the speaker is *probably* making these assumptions, but it is still *logically possible* that they do *not* think that the proof will persuade you. They *likely* assume that it is a way to change your belief, or they would be irrational to waste their time offering you such a proof. But assumptions that are necessary to retain one's rationality in *making* a statement are *not* the same as assumptions that are *logically implied* by the *content* of the

statement one makes. (See Paul Grice's distinction between mere *conversational implication* and *statement implication* [1975].) Fortunately, this item was not reused on Form E.

5.3 Deduction

This section assesses one's ability to recognize whether or not the truth of given premises entails the truth of proposed conclusions. This time the test's directions are perfectly clear:

> If you think [the suggested conclusion] *necessarily* follows from the statements given, [answer] "CONCLUSION FOLLOWS"…. If you think it is *not a necessary conclusion* from the statements given, [answer] "CONCLUSION DOES NOT FOLLOW," even though you may believe it to be true from your general knowledge.

All 16 items in this section on Form B concern five enthymemes—two premises of a syllogism are provided, followed by suggested conclusions. On Form A, two of the five target arguments are better characterized as propositional arguments. This primary focus on logical entailment, especially on syllogistic arguments, seems a bit excessive to Fisher and Scriven's taste (1997, p. 119), but not to mine—it is amazing how much of our everyday reasoning is deductive and represented syllogistically. However, I find that there is a complication involving one of the enthymemes on Form A and two of them on Form B. This issue concerns what is called "existential import."

The tests' example will help to illustrate: We are told that some holidays are rainy and that all rainy days are boring. On the basis of these two premises, it is said to follow that some holidays are boring. As the directions for this section tell us, "'Some holidays are rainy' means *at least* one, possibly more than one, and *perhaps* even all holidays are rainy." So, when one claims that *some* Xs are (or are not) Ys, *one is going on record as committing oneself to the existence of at least one* X. That same "existential commitment" is *not* necessarily made when we make *universal* claims; e.g., we could truthfully claim that *all* our job applicants must take the W-G (it's company policy!), even when we have no such applicants. And therein lies the issue with our three problematic enthymemes.

For example, on Form A, we are told that all members of the orchestra enjoy playing classical music and that they all spend long hours practicing. Does it follow that *some* musicians who spend long hours practicing enjoy classical music? Notice that we have two universal premises, and yet this would be a conclusion that commits us to the existence of at least one musician who spends long hours practicing and enjoying classical music. The *premises* have not technically established the existence of such a musician, and the directions tell us to determine what follows "from the statements given." So *technically* this conclusion does *not* follow.

But being able to catch an argument on a technicality and being a charitable critical thinker are two different things. And I think that is the case here. The question to ask of *any* enthymeme under discussion is, *Does this conclusion follow given the obvious existence of at least one member in the categories discussed in the premises?* For example, would you grant the speaker the existence of at least one member of the orchestra? The person stating the enthymeme obviously thinks at least one such orchestra member exists (otherwise that person wouldn't likely be talking about how hard they're practicing!). And now the conclusion indeed *does* follow, *assuming that there exists an orchestra member*. Rather than catching the argument on a technicality, we are acknowledging that the argument, *like virtually every argument*, also has an *implicit* premise that is perfectly acceptable.

The willingness to judge that a conclusion follows, given reasonable existential commitments, is especially important in one of the exercises on Form B (omitted on Form E): We are told that all radicals are members of a political minority and that no patriot is a radical. Does it follow that *some* members of a political minority are unpatriotic? Well, yes, but *only if* there exists at least one radical, which is a perfectly acceptable *implicit* premise.

The irony here is that someone who has taken formal logic hyperseriously, so as to hone their CT skills with respect to deductive arguments, might easily suffer a lower score on the W-G as a result. That should not happen. And it could easily be avoided by stating in the directions that test subjects are allowed to make reasonable assumptions regarding the existence of objects discussed in the premises, when deciding whether the conclusion deductively follows or not.

Form A also has one item that is just plain incorrectly scored: From premises stating that a school system has 52

classes, with each class containing 10-40 pupils, it is said to follow that there are at least 550 pupils. Clearly, all that follows is that there are at least 520 pupils. Fortunately, this item was not reused on either the Short Form or Form D.

5.4 Interpretation

This is an oddly named section, in that it suggests that the task will be one of reading comprehension more than anything else. But the real task in this section is to correctly judge the strength with which the truth of premises indicates the truth of a conclusion. Once again, the directions are a source of confusion:

> Each exercise below consists of a short paragraph followed by several suggested conclusions.
>
> *For the purpose of this test, assume that everything in the short paragraph is true.* The problem is to judge whether or not each of the proposed conclusions logically follows beyond a reasonable doubt *from the information given in the paragraph.*
>
> If you think that the proposed conclusion follows beyond a reasonable doubt (even though it may not follow absolutely and necessarily), then [answer] "CONCLUSION FOLLOWS"…. If you think the conclusion does *not* follow beyond a reasonable doubt from the facts given, then [answer] "CONCLUSION DOES NOT FOLLOW." Remember to judge each conclusion independently.

The confusion found in these directions is *not* the one manufactured by McPeck (1981, pp. 138-140) and Fisher and Scriven (1997, p. 120), who think that the directions give test subjects permission to base their judgments on whatever they simply believe to be true beyond a reasonable doubt. As a result, McPeck thinks that test subjects could "justifiably regard any or none of these inferences as following 'beyond a reasonable doubt from the paragraph'." Not so, on both accounts—the directions here explicitly state that test subjects are to base their judgments on "*the information given in the paragraph*," period.

The real problem is one that we've seen before: The directions first instruct you to judge whether the conclusion "logically follows" from the information provided. This suggests that the arguments are to be treated as *deductive* arguments and that the standard of assessment should be logical *entailment*, i.e., that there is no *logical* possibility that the conclusion is false while the premises are (assumed) true. But then, in the same breath, you are instructed to judge whether the conclusion is "beyond a reasonable doubt" based on the provided information. This suggests that the arguments are to be treated as *inductive* arguments instead, assessed in terms of whether or not there is a *real* possibility that the conclusion is false while the premises are (assumed) true. Once again, the ambiguity is resolved by a parenthetical remark, in the third paragraph of the directions: "[the conclusion] need not follow absolutely and necessarily." Therefore, the arguments are to be treated as *inductive* arguments. Missing that embedded remark, however, would have a significant effect on one's answers. That this could very well happen is illustrated by no less a critical-thinking expert as Trudy Govier (1987, p. 256), who concluded that the W-G fails to address inductive reasoning (when, in fact it does, in this section, the first, and the last).

The direction's example helps to clarify that the question at issue is whether the premises *inductively* support the conclusion "beyond a reasonable doubt": "None of the children in this study had learned to talk by the age of six months" is said to follow from the stated fact that "the size of the spoken vocabulary at eight months was zero words." While it is *logically* possible (i.e., it's not a contradiction) that a child in the study *had* learned to talk at six months and simply made no utterance during the study, it is *not* a *plausible* enough possibility to constitute reasonable grounds for doubt.

You might be wondering by now what constitutes a plausible enough possibility so as to be grounds for reasonable doubt. There is no exact answer to that question, because it is a vague concept; i.e., there is no defining set of necessary and sufficient conditions for when a conclusion is "beyond a reasonable doubt" based on assumed premises. This does not mean that the concept is subjective or worthless; it just means that you must apply the concept to clear-cut cases if you want uncontroversially right or wrong answers. This is usually achieved by the items in this section of the test. But occasionally there is room for reasonable disagreement as to whether or not the evidence provided in the premises makes

doubt about the conclusion reasonable or not. For instance, in one of the scenarios, someone who is normally a good sleeper has trouble getting to sleep whenever they drink coffee in the evening, which they do about twice a month. Is it beyond a reasonable doubt that they better not drink coffee when they want to fall asleep promptly at night? We are told that it is; but is that correct? There is no mention of how long this person has done this informal study. If only a month or so, they have fairly weak evidence—their sample is too small. If this has been going on for years, then it would be reasonable to say their evidence justifies that recommendation beyond a reasonable doubt—even if it's not the caffeine that is the cause of their sleeplessness but just a placebo effect.

I should also mention that there is an inconsistency between two answers in this section on Form B: It is scored as beyond a reasonable doubt that "No ordinary form of communication could account for the occurrence of the wife's dream and the husband's accident at the same time," and yet it is scored as *not* beyond a reasonable doubt that "The dream was a chance coincidence that was not really influenced by the accident." But, the degree to which there was *no* form of communication between the couple *is* the degree to which the wife's dream was a chance coincidence *not* influenced by the husband's accident. This inconsistency is not excused by the instruction to "judge each conclusion independently." And nothing was stated in the scenario that would make a supernatural form of influence a reasonable possibility. Fortunately, this item is not reused on Form E.

5.5 Evaluation of arguments

The function of this section is to assess one's ability to correctly judge the strength with which the assumed truth of the premises indicates the truth of the conclusion. Once again, the directions are unclear and distracting:

> *For an argument to be strong, it must be both important and directly related to the question*.
>
> An argument is weak if it is not directly related to the question (even though it may be of great general importance), or if it is of minor importance, or if it is related only to trivial aspects of the question.

> Below is a series of questions. Each question is followed by several arguments. *For the purpose of this test, you are to regard each argument as true*. The problem then is to decide whether it is a *strong* or a *weak* argument….

This characterization of what makes an argument strong or weak is of no help: What it is for an argument to be "important" and "directly related" to "the question" is mysterious. And again, *arguments* are neither true nor false; *statements* can be true or false. Arguments are either valid or invalid, cogent or not cogent, strong or weak. The directions should instead stipulate that the *premises* in the proposed arguments are to be regarded as true. The task, then, is to judge whether or not the premise provides a good reason to think that the proposed conclusion is true, if it provides any reason at all.

What constitutes "good reason" sufficient to make an argument "strong" is vague—which, again, is *not* to say that the cogency of an argument is purely subjective. It merely implies that one needs to craft items on the test that have clear-cut answers, which is not always achieved. For example, in one item, an argument concludes that having a strong labor party does not promote the general welfare of people in the U.S., in light of the strikes unions have called in a number of important industries. This is scored as a weak argument. But there can be reasonable differences of opinion about the strength of this argument, depending on one's background beliefs about the overall impact of industrial labor strikes compared to the overall beneficial effects of unions.

At the end of the directions for this section, the W-G includes the following sentence: "When the word 'should' is used as the first word in any of the following questions, its meaning is, 'Would the proposed action promote the general welfare of the people in the United States?'" This instruction is problematic for two reasons (setting aside how odd this requirement must strike test subjects in other countries). First, it can be quite disruptive for test subjects who take the stipulation *literally* and try to *substitute* the suggested phrase for the word 'should' in the test items. Second, adopting the required utilitarian point of view narrows the criteria by which one is to judge the strength of the arguments. This can drastically affect one's answers. For example, that an action is unconstitutional or violates an individual's rights might strike one as a strong

reason against performing that action, but they become weak (i.e., irrelevant) reasons, if one is forced to judge *solely* in light of whether the action will promote the general welfare for the U.S. population. This issue plagues, for instance, three items in this section of Form B:

- That the government would be able to control inflation, which is seriously threatening to bring about economic depression, is scored as a strong reason for why the U.S. government should nationalize all major industries. But, if one believed that there were *other* means of controlling inflation that did not interfere with liberties to this extent, one would reasonably construe this as a weak argument.

- That the government's money is the taxpayers', who are already taxed too heavily, is scored as a weak reason against the government's subsidizing farmers for their soil conservation practices. But, someone who believed that negative rights take priority in moral reasoning would reasonably judge this to be a strong reason against this government program. And, even using solely a utilitarian calculus as directed, it is *not* obvious that taxation for the purpose of such a subsidy program maximizes the general welfare, *especially* given the premise that the tax burden is too heavy already.

- That a lowering of air and water standards will inevitably lead to loss of human life is scored as a strong reason to maintain those standards despite their causing higher consumer prices for electricity and manufacturing. But this argument is weakened by virtue of a *reductio ad absurdum* criticism: Driving cars, riding bicycles, and merely getting out of bed in the morning "inevitably lead to loss of human life" too. Therefore, having *that* consequence, in and of itself, can't be adequate grounds for objecting to a lowering of air and water standards. This perfectly good criticism against the argument would be *verboten*, however, given the utilitarian restriction set down in the directions. And, once again, the utilitarian calculus itself doesn't provide such a clear answer: how much loss of life, compared to how much loss of life due

> to an increase in these consumer prices? One could reasonably believe that the latter is more lethal.

Other critics worry instead that test subjects will too drastically stray from the required criterion of whether the actions in question "promote the general welfare of the people of the United States" and judge the strength of the arguments merely in light of "political or ideological commitments" (Fisher and Scriven, 1997, p. 121). "Political liberals and conservatives might well disagree on every one of these items" (McPeck, 1981, p. 141). These qualms have long been expressed (Ennis, 1958, p. 157), and they are legitimate concerns—items that so strongly tempt test subjects to stray from the assigned task diminish this section's construct validity.

The current publishers of the W-G would seem to disagree, blaming the test subjects instead of the test: Test subjects who score poorly on this section are informed, in their "Profile Report," that their "score suggests below average skill and consistency when this individual needs to: evaluate arguments based on the relevance and strength of the evidence supporting them [and] analyze information objectively, without allowing preferences or emotions to influence evaluations" (Pearson Education, 2014). For reasons just discussed, I would strongly question anyone who concludes that a candidate being considered for hiring or promotion is either biased or emotional in their judgments on the basis of such a low score.

6. Conclusion

The W-G, in its many manifestations, has had many serious flaws, most of which remain in currently used versions, because those flaws involve the tests' misleading directions, which have not been altered since at least 1980. Whether these problems have simply gone unnoticed for decades, or were willingly ignored, I cannot say. I can say, however, that they severely diminish the tests' validity, they are unnecessary and easily remedied, and they should be remedied as soon as possible. The irony is that the more test subjects know about formal and informal logic, the worse they might well do on the W-G; and that's *not* how things should work for a CT assessment test.

World-wide, professional hiring and development services place a great deal of trust in the W-G (e.g., PDI Ninth

House). It's time for the W-G to do a better job of earning that trust.

Acknowledgements: I thank my anonymous referees for their many wonderful suggestions; I hope I've done them justice.

References

Ejiogu, K., Yang, Z., Trent, J., & Rose, M. (2012.) "Understanding the Relationship Between Critical Thinking and Job Performance." Retrieved from http://us.talentlens.com/wp-content/uploads/TalentLens-CriticalThinking-and-Performance.pdf

Ennis, R. (1958). "An Appraisal of the Watson-Glaser Critical Thinking Appraisal." *Journal of Educational Research*, 52(4), 155-158.

Ennis, R. (2013). "Critical Thinking Across the Curriculum: The Wisdom CTAC Program." *INQUIRY: Critical Thinking Across the Disciplines*, 28(2), 25-45.

Fawkes, D., Adajian, T., Flage, D., Hoeltzel, S., Knorpp, B., O'Meara, B., & Weber, D. (2001). "Examining the Exam: A Critical Look at the Watson-Glaser Critical Thinking Appraisal Exam." *INQUIRY: Critical Thinking Across the Disciplines*, 20(4), 19-33.

Fisher, A. & Scriven, M. (1997). *Critical Thinking: Its Definition and Assessment*. Norwich, UK: Centre for Research in Critical Thinking.

Glaser, E. (1941). *An Experiment In the Development of Critical Thinking*. New York, NY: Bureau of Publications, Teachers' College.

Govier, T. (1987). *Problems in Argument Analysis and Evaluation*. Dordrecht: Foris.

Grice, P. (1975). "Logic and Conversation." In P. Cole & J. Morgan (Eds.), *Syntax and Semantics: Speech Acts* (pp. 41-58). New York, NY: Academic Press.

Hunt Executive Search. (2013). http://www.huntgroup.com/Critical-Thinking.asp

Lawshe, C. (1975.) "A Quantitative Approach to Content Validity." *Personnel Psychology,* 28, 563-575.

McPeck, J. (1981). *Critical Thinking and Education*. New York, NY: St Martin's Press.

McPeck, J. (1984). "The Evaluation of Critical Thinking Programs: Dangers and Dogmas." *Informal Logic*, 6(2), 9-13.

Menkes, J. (2005). "Hiring for Smarts." *Harvard Business Review*, 83, November, 1-10.

Norris, S. & Ennis, R. (1989). *Evaluating Critical Thinking*. Pacific Grove, CA: Midwest Publications Critical Thinking Press.

Pearson Education. (2012). *Watson-Glaser Critical Thinking Appraisal User-Guide and Technical Manual*. Retrieved from http://www.talentlens.co.uk/assets/news-and-events/watson-glaser-user-guide-and-technical-manual.pdf

Pearson Education. (2014). *Watson-Glaser II Critical Thinking Appraisal Profile Report*. Retrieved from http://www.thinkwatson.com/assessments/watson-glaser/profile-report

Pearson TalentLens. (2011). *Watson-Glaser Critical Thinking Appraisal Unsupervised.* Pearson TalentLens.

Pearson TalentLens. (2013). *The Bar Course Aptitude Test*. Pearson TalentLens.

Possin, K. (2008). "A Field Guide to Critical-Thinking Assessment." *Teaching Philosophy*, 31(3), 201-28.

Ryan, A. M. & Sackett, P. (1987). "A Survey of Individual Assessment Practices by I/O Psychologists." *Personnel Psychology*, 40, 455-488.

Watson, G. & Glaser, E. (1980). *Watson-Glaser Critical Thinking Appraisal Forms A and B*. San Antonio, TX: The Psychological Corporation.

Watson, G. & Glaser, E. (1994). *Watson-Glaser Critical Thinking Appraisal Short Form*. Harcourt.

Watson, G. & Glaser, E. (2009). *Watson-Glaser Critical Thinking Appraisal Forms D and E*. Pearson.