Robust Evidence and Secure Evidence Claims*

Kent W. Staley†‡

Many philosophers have claimed that evidence for a theory is better when multiple independent tests yield the same result, i.e., when experimental results are *robust*. Little has been said about the grounds on which such a claim rests, however. The present essay presents an analysis of the evidential value of robustness that rests on the fallibility of assumptions about the reliability of testing procedures and a distinction between the strength of evidence and the security of an evidence claim. Robustness can enhance the security of an evidence claim either by providing what I call second-order evidence, or by providing back-up evidence for a hypothesis.

1. Introduction. Many philosophers (Whewell 1989, 138–160; Peirce 1992, 29, 138; Wimsatt 1981; Hacking 1983; Trout 1993; Culp 1994, 1995) have claimed that evidence for a theory is better when multiple independent tests yield the same (convergent) positive result, i.e., when experimental results are *robust*. Robert Hudson has recently denied this claim (Hudson 1999). The evidential value of robust evidence has been argued for from a Bayesian perspective (Franklin and Howson 1984). This essay seeks to demonstrate the evidential value of robustness from a non-Bayesian point of view and to identify some limits to the evidential value of robustness.

The present analysis also draws attention to two distinctions that writers on evidence do not typically draw. First, I distinguish between first-order and second-order evidence in order to shed light on the use of data to

*Received June 2003; revised January 2004.

†To contact the author write to Department of Philosophy, Saint Louis University, 3800 Lindell Blvd., St. Louis, MO 63108; email: staleykw@slu.edu.

‡Audiences to presentations at Saint Louis University, the Northwest Philosophy Conference, and the Twelfth International Congress of Logic, Methodology, and Philosophy of Science have assisted the author in refining this analysis. I have had helpful discussions with George Terzis, James Marcum, Wayne Myrvold, and Bill Harper. Robert Hudson and an anonymous referee for another journal helped steer me toward an improved formulation, and two anonymous referees for this journal provided helpful commentary and suggestions.

Philosophy of Science, 71 (October 2004) pp. 467–488. 0031-8248/2004/7104-0003$10.00

provide evidence for an evidence claim. I also distinguish between the strength of evidence and the security of an evidence claim. The strength of evidence concerns how strongly data indicate the correctness of a hypothesis. The security of an evidence claim concerns the degree to which that claim is susceptible to defeat from the failure of an auxiliary assumption.

Putting aside the vexing question of whether evidence is made stronger by being robust, I argue that robustness does enhance security, although this can happen in two ways: (1) An evidence claim based on the results of one test can be made more secure against being *wrong in extent* (the evidence is weaker than claimed) by appeal to convergent results from an independent test. Such use of robustness illustrates also the use of second-order evidence: the second result is used to provide evidence that the first result is evidence of the claimed strength for the hypothesis of primary interest. (2) An evidence claim based on the combined results of independent sources can be made more secure against being *categorically wrong* (the results in fact do not provide evidence at all). In such a case, robust results are combined to make a stronger evidence claim than could be made from the result from any one source. The fact that the results come from independent sources is of no help here in avoiding being wrong in extent, since the claimed strength of evidence requires the correctness of all assumptions underlying the different sources.

For clarity, my analysis employs Deborah Mayo's error-statistical theory of evidence. However, my central claims do not depend on the specifics of that theory, but will hold for any theory of evidence in which evidential relations supervene on facts about the reliability of testing or inferential procedures. I illustrate these points with an example from the recent history of experimental particle physics.

My procedure will be as follows: In Section 2, I give a brief outline of the error-statistical theory of evidence. Sections 3 and 4 explain, respectively, the distinction between first- and second-order evidence and the distinction between strength and security of evidence. In Section 5, I employ these distinctions to give a general analysis of two uses of robust evidence. I illustrate this analysis in Section 6 with a discussion of some of the uses of data in the argument presenting the first evidence for the top quark. Section 7 summarizes my central claims.

2. The Error-Statistical Theory of Evidence. On the error statistical account, an experimental result E counts as evidence for a hypothesis H only if H passes a severe test with E, where H passes a severe test T with outcome E just in case E fits H, and the probability of H passing T with an outcome such as E (i.e., one that fits H as well as E does), given that H is false, is very low (Mayo 1996, esp. 178–187).

The central idea behind the error-statistical approach is that only test results that discriminate one hypothesis from its alternatives count as evidence for that hypothesis, where discrimination is a matter of indicating reliably the truth of that hypothesis as opposed to its alternatives. The severity requirement thus must be contrasted with the requirement of merely fitting the hypothesis (either probabilistically or as a logical consequence). The severe test requirement ensures that test results constituting evidence for a particular hypothesis are of a kind that would be highly improbable were it the case that one of the alternatives was true. In this assessment it is important to note that all statistically relevant aspects of the testing procedure must be taken into account. However, the assessment may be, and often is, made informally, without invoking a precise quantitative probability model (see Mayo 1996, 64).

3. First- and Second-Order Evidence. The error-statistical theory conceptualizes evidence as an empirical matter. Whether E is evidence for H cannot be decided on the basis simply of analysis of sentences expressing E and H and formal relations between them. Whether an evidential relation obtains is to be determined empirically (see also Achinstein 1995, 2001). Given an empirical concept of evidence, it is useful to distinguish between *first-order evidence* and *second-order evidence*: If some fact E constitutes first-order evidence with respect to a hypothesis H, then it provides some reason to believe (or indicates) that H is the case. If a fact E is second-order evidence with respect to a hypothesis H, then it provides some reason to believe (or indicates) that some distinct fact E' is first-order evidence with respect to H.

The distinction between first- and second-order evidence is here relativized to a particular hypothesis of interest. Given that a hypothesis is of primary interest to us, data might function either as evidence for that hypothesis (first-order), or as evidence that some other data are evidence for that hypothesis (second-order). The distinction is thus not an ontological dichotomy between different kinds of facts. The same fact might, in different contexts, function either as first- or second-order evidence with respect to the same hypothesis, provided that it satisfies both the requirements for being evidence for that hypothesis, and for being evidence for the claim that some other fact is evidence for that hypothesis.

Often, when estimation of the strength of evidence employs a probability model, second-order evidence derives from the testing of that model. Such tests may yield either second-order evidence with respect to the primary hypothesis in question, or simply first-order evidence in support of one or more of the assumptions on which the probability model is based. It is important to note that the latter does not automatically translate into evidence in support of a first-order evidence claim for the primary

hypothesis. Suppose that E is evidence for an assumption on which a first-order evidence claim C ("E' is evidence for H") is predicated. On the error-statistical account, E will constitute second-order evidence for H only if E is also the outcome of a test that is severe with respect to the hypothesis C. (In many cases, including some of the examples discussed below, it may be unclear whether data is being used to provide second-order evidence for the primary hypothesis or simply evidence for an auxiliary assumption. The important point is that robust results can be used for either purpose.)

4. Strength of Evidence and Security of Evidence Claims. Although the evidence concept can be used categorically, scientists often attempt to characterize how strongly data indicate the correctness of a hypothesis. Such estimations of strength of evidence are often qualitative, using terms such as "suggestive," "persuasive," "weak," "strong," "compelling," or "conclusive." Philosophers of science (e.g., Carnap 1962) have long sought to quantify this dimension of evidential assessment. Such efforts aside, reports of experimental outcomes do often involve quantitative measures deriving from specific statistical tools such as significance tests, chi-squared fits, likelihood ratios, and the like. Although these classical statistical measures do not yield the kind of confirmation measure sought by philosophers, they do sometimes function as partial quantitative indicators of evidential strength (see Staley 2004, ch. 6).

Another aspect of evidential assessment has received less attention from philosophers: the *security* of an evidence claim. The security of an evidence claim concerns the degree to which the claim that some result is evidence for a hypothesis is itself susceptible to defeat from the failure of an auxiliary assumption.

When the strength of an evidence claim is evaluated (whether categorically, qualitatively, or quantitatively) the assumptions on which the evidence claim relies are *used* in that evaluation, and are not simultaneously the subject of evaluation. This does not, however, mean that confidence in such assumptions is absolute. Even if such confidence were absolute, it might be misplaced. Security is relevant because of the fallibility of auxiliary assumptions.[1] Thus one can, as a separate step in the assessment of evidence, ask two kinds of question about such an assumption:

1. "Auxiliary assumption" is here used as a generic term for assumptions that could potentially defeat an evidence claim, regardless of the precise role played by that assumption in supporting the evidence claim. As Giora Hon has argued, the epistemology of experiment can be fruitfully analyzed through a typology of sources of error (Hon 2003). Hon distinguishes between the background theory of the experiment, assumptions about the apparatus employed, the observation or recording of data, and

1. How strong is the evidence in support of the assumption employed?
2. How sensitive is the primary evidence claim to the failure of the assumption?

Thus the security of an evidence claim can be enhanced either by resting upon assumptions for which the evidence is exceptionally strong, or by resting upon its assumptions rather lightly, so to speak. No doubt scientists will always prefer to use only those assumptions for which they have good evidence. Even when the evidence for an auxiliary assumption is judged to be exceptionally strong, however, the latter security consideration matters, because such judgments can themselves be mistaken.

It is not obvious whether robustness by itself *strengthens* evidence. Suppose that two sets of test results are statistically equivalent with respect to some hypothesis (e.g., both tests yield results that reject the null hypothesis at the same significance level, both tests have the same power, they involve samples of the same size, etc.) but one set combines data from several different experiments using different assumptions, while the other draws upon only one experiment relying on a single, smaller set of assumptions. Is the former, ceteris paribus, stronger evidence for the hypothesis than the latter? I have not found a satisfactory resolution to this question, and so I set it aside.[2] Instead, I intend to argue that the robust result is, under certain circumstances, more *secure* than the non-robust result. Furthermore, I will show that such enhancement of security can be obtained by using robustness to address either of the two types of questions mentioned above, and that robustness can be used in different ways to achieve security against at least two kinds of error.

5. Two Uses of Robust Evidence. Suppose that a report of an experimental outcome claims both that the results of the experiment at hand constitute evidence for a particular hypothesis, and that the evidence is of some specified degree of strength. Such a claim can be wrong in two ways, and convergent results can be used to protect against either kind of error:

---

the interpretation of results. Evidence claims rest on assumptions concerning all four aspects of experiment, and can be defeated by the failure of an assumption of any of these kinds.

2. Treating the combination of different test outcomes as the outcome of a single test is a task for metaanalysis. As an anonymous referee for this journal pointed out, such metaanalysis yields valid statistical assessments only if the assumptions involved in *each* of the combined tests are satisfied. My point is that, supposing this to be the case, it remains unclear whether such a combined result would more *strongly* indicate the correctness of the hypothesis than a statistically equivalent result of a single test.

A. An evidence claim based on one test can be made more secure against being wrong in extent (the evidence is weaker than claimed).
B. An evidence claim based on the combined results of independent tests (i.e., tests based on independent assumptions) can be made more secure against being categorically wrong (the results in fact do not provide evidence at all).

5.1. Securing the Degree of Strength. Use (A) can take two forms, corresponding to questions (1) and (2) listed in Section 4. Suppose that one has data from two independent tests, T_1 and T_2, such that both tests yield the same result. One might then cite results from T_2 as evidence for the assumptions underlying an evidence claim based on results from T_1. Such use of robustness may also amount to second-order evidence if the second result is used to provide evidence that the first result really is evidence of the claimed strength for the hypothesis of primary interest. Here convergent results are cited to address questions of type (1).

This strategy rests on the argument that it is highly improbable that T_2 would produce results in agreement with results from T_1, if the underlying assumptions of the evidence claim based on T_1 were false. (This is an application of the severe test requirement.) The strategy only works when such an argument can be sustained, and it will sometimes fail. The argument fails when convergence is likely to occur regardless of the correctness of the assumptions in question. Here again there are two possibilities of interest.

Such an appeal to second-order evidence can fail when one test is likely to produce the result in question regardless of whether the assumption in question is true. I will call this *spurious convergence*. Consider two particle detectors arranged as coincidence indicators, so that a particle passing through one will almost certainly pass through the other, producing two nearly simultaneous signals. Assume that the two detectors are based on entirely different technologies and rely on different physical principles, so as to constitute independent means of detection, and that both detectors produce a signal at about the same time. The results satisfy the robustness requirement, being both convergent and produced independently. If, however, the second detector were so noisy that it had a 50% chance of producing a signal in the absence of any particle, we could safely conclude that the convergence of these independently produced results is without evidential value. More specifically, the fact that a signal was generated by the second detector does not help support the assumption that the first detector is reliable, since given the noisiness of the second detector and the occurrence of a signal from the first detector, whether the first detector

is reliable does not affect the probability of a signal from the second detector.

It is worth noting that such an example, though contrived, shows that convergent results are not guaranteed to have evidential value, even with respect to the security of an evidence claim.

Another way in which the appeal to second-order evidence can fail is that *apparently* robust results might in fact derive from non-independent tests (or, independence being a matter of degree, tests that are less independent than assumed). These constitute *failures of independence*. Suppose once again that two detectors are arranged in coincidence in order to search for evidence of a specific particle, Π. Shielding is used to eliminate background that might otherwise produce coincident signals of the sort sought. Here again we might be mistaken in using the coincidence between signals as evidence for the reliability of one detector for detecting Π particles if some source of background is able, unbeknownst to us, to penetrate the shielding. We would then be assuming falsely that a coincidence between signals would be highly improbable if it were not true that one of the detectors was a reliable discriminator between Πs and other particles. The presence of background means that such coincidences will occur with high probability even if neither detector reliably detects Πs. Here our problem is that, although tests based on data from the two detectors rest on different assumptions about the apparatus, they both rely on a single assumption about the shielding, the failure of which defeats the reliability of both detectors simultaneously.

A further requirement can prevent both modes of failure. In addition to robustness of evidence, or *convergent validation* of the results, we can require that the results provide *discriminant validation*. Discriminant validation requires that the different sources of evidence do not yield convergent results when the phenomenon to be detected or measured is absent.

The literature on robust evidence in philosophy of science has tended to neglect discriminant validation, although it is prominent in the early writings on the topic. In his influential work on robustness, William Wimsatt (1981) draws upon work by Donald Campbell and Donald Fiske (Campbell and Fiske 1959). Campbell and Fiske describe convergent validation as "confirmation by independent measurement procedures." But Campbell and Fiske also note that "[f]or the justification of novel trait measures, for the validation of test interpretation, or for the establishment of construct validity, discriminant validation as well as convergent validation is required" (Campbell and Fiske 1959, 81). Discriminant validation is a process of checking to see whether a particular process produces results that correlate too highly with the results of processes that should yield uncorrelated results.

From an error-statistical perspective, the relevance of discriminant va-

lidity can be explained: agreement of results from independent processes is evidence for the reliability of any one of those processes (or for any evidence claim based on the results of that process), only if such agreement amounts to a severe test of the reliability of that process (or of the evidence claim based on its results). Mere agreement does not suffice. One should be able to argue that the agreement is of a sort that would be very improbable if the reliability assumption (or evidence claim) in question is not true. Discriminant validation helps to establish that this severe test requirement is met by showing the test to be sensitive to the kinds of facts being claimed.

Wimsatt notes (1981, 156-59) the importance of discriminant validation for purposes of identifying failures of independence, though not for identifying spurious convergence. (This is not unreasonable. Real-life examples of spurious convergence seem much harder to find than failures of independence.) One aim of this paper is to restore discriminant validation to its rightful place alongside robustness considerations.

Consider how both of the examples of failed robustness arguments succumb to the test of discriminant validation. In the case of spurious convergence, the results meet the requirements of convergent validation, but fail the test of discriminant validation. The second detector would frequently deliver such a confirming signal even if we employed it as an anti-coincidence detector. In the case of failure of independence, we might arrange to ensure that no Πs reach the two detectors, and then look to see how often the two detectors deliver a coincident signal. The neglected background would continue to deliver coincident signals at a higher rate than expected, although the phenomenon sought is absent.

Thus far I have been considering the use of convergent results to serve the purpose of addressing questions of type (1) as discussed in Section 4: the results of only one test are considered as first-order evidence for the primary hypothesis; those of the other serve as evidential support for assumptions about the first test on which that evidence claim rests. But one could also use convergent results from a second test to serve as a kind of "back up" evidence against the possibility that some assumption underlying the first test should prove false.

The difference is similar to the following: An engineer has a certain amount of material with which to construct the pilings for a bridge. Calculations show that only 60% of the material is needed to build a set of pilings sufficient to meet the design specifications, but the extra material, if not used, will simply go to waste. The engineer decides to "over-engineer" the pilings with the extra material. Two possibilities are to use 60% of the material to produce a single set of pilings, and use the extra material to reinforce those, or to use the extra material to produce ad-

ditional pilings. The former case is analogous to the use of convergent results to support auxiliary assumptions (question 1).

Like the engineer who chooses to build extra pilings, the scientist might use convergent results to address question (2) in Section 4: by showing that one has a kind of back-up source of evidence that rests on different assumptions than those behind the primary evidence claim, one might be protected against the failure due to a wrong assumption of one's claim about how strong the evidence is for a hypothesis. In effect, this is to claim that, although one's assumptions might be wrong, one's claim that the hypothesis has evidence of some specified strength in support of it would still be correct (though not for the reasons initially given).

5.2. Securing Categorical Evidence Claims. Suppose that one wishes to make as strong an evidence claim as the data permit, and is willing to risk being wrong about the extent of that evidence. In that case one may wish to take convergent results from independent tests T_1 and T_2 and combine them into a single result supporting a strong evidence claim. This decision may be made more reasonable by having strong evidence supporting the assumptions underlying the use of data from T_1 and T_2. However, in this case the fact that results from T_1 and T_2 converge cannot be used as such supporting evidence for all such assumptions without creating a vicious circle. Any appeal to the results of T_1, for example, will make *use* of a subset of the overall assumptions. The need to support all relevant assumptions in a non-circular manner thus limits the potential for using the robustness of one's results to address question (1) in Section 4.

Nonetheless, the security of this evidence claim can be enhanced by use (B) of the convergent results: Here one is concerned to avoid the error of claiming some evidence when there is none. If one takes the convergent results from independent tests and combines them into a single evidence claim, the robustness of the combined result enhances the security of that single claim by ensuring that the failure of a single, non-shared assumption will only invalidate part of the results underlying one's evidence claim. Provided that the remaining results suffice to constitute *some* evidence, one will at least not be wrong about having some evidence in support of the primary hypothesis, although one might be wrong about its strength.[3]

3. Note that on the error statistical account, evidence is a threshold concept. Although no precise degree of severity is specified as necessary for evidence, results that only qualify a hypothesis as having passed a test with minimal severity will not count as evidence. Peter Achinstein (2001) has argued independently that evidence is a threshold concept. Accordingly, there will be cases of robust evidence that fail to meet the requirement that defeat of one assumption not shared by all relevant tests leaves intact a valid though weaker evidence claim.

In such uses robust results are combined to make a stronger evidence claim than could be made from the results of just one test. The fact that the results come from independent tests is of no help here in avoiding being wrong in extent, since the claimed strength of evidence requires the correctness of all assumptions underlying the different tests. Indeed, the reliance on a larger number of distinct assumptions when combining results in this way (e.g., through metaanalysis) *undermines* security with respect to the claimed strength of evidence, as it opens up new possibilities for error relative to the use of results from just one test. In other words, uses (A) and (B) of robust results are incompatible — one must choose.

6. Robustness and Security in the Evidence for the Top Quark. Physicists' Standard Model postulates six "flavors" of quarks. In April 1994, the Collider Detector at Fermilab (CDF) collaboration, based at Fermi National Accelerator Laboratory, announced that they had found "evidence" for the existence of the top quark (Abe et al. 1994), the last of the six flavors to be experimentally confirmed. In the present discussion I will focus on the analysis CDF employed in justifying this claim.

A thorough discussion of CDF's analysis would go beyond the constraints of the present essay (see Staley 2004). Here I will sketch CDF's general approach in searching for evidence of the top quark, and highlight the deployment of robustness considerations in some aspects of the argument.

CDF employed a detector surrounding a collision point on Fermilab's proton-antiproton colliding accelerator (the "Tevatron"). Collisions between protons and antiprotons that have been accelerated to nearly the speed of light release enormous amounts of energy. If the top quark exists, then such collisions can result in the creation of a top quark–anti-top quark ($t\bar{t}$) pair. According to the Standard Model, the $t\bar{t}$ pair thus produced should have certain characteristic decay modes. Detection of the top quark would occur through detection of these "signatures." In particular, each top quark would nearly always decay into a W boson and a b ("bottom") quark—the t into a W^+ and a b, the $\bar{t}$ into a W^- and a $\bar{b}$. CDF attempted to identify top quark events by detecting the decay products of the W boson and b quark.

Two decay modes of the W bosons formed the basis for CDF's search. In the first, both W bosons decay into a lepton (either an electron e or muon μ) and its associated neutrino (ν_e or ν_μ). These were known as "dilepton" decays. In "lepton plus jets" decays, one of the Ws decays to a lepton-neutrino pair, and the other decays into a quark-antiquark pair, yielding two narrow "jets" of quark-bearing hadrons.

Identifying events with these characteristics relied on a set of *cuts* specifying the measured characteristics of an event that would qualify it as a

top quark *candidate event*. Once the cuts had been chosen, an algorithm could be written that would scrutinize each event in the data set and determine whether it was a top quark candidate. For any given set of cuts, some non–top quark (background) events would pass. A well-chosen set of cuts, however, would enable CDF to detect the existence of the top quark by means of a significant excess of candidate events beyond the expected background. Such a procedure is known as a *counting experiment*.

CDF employed three different counting experiments, applied to a set of data collected during 1992–93 ("run Ia").[4] The *dilepton* counting experiment looked for events yielding a pair of energetic leptons, at least two energetic jets, and a neutrino. (Since neutrinos interact very weakly with matter, the presence of a neutrino is inferred when, on the basis of the measured energies of other decay products, conservation of energy considerations indicate that a significant amount of "missing" energy was carried off by an undetected particle.) When applied to the run Ia data set (about 16 million events), the dilepton algorithm identified two candidate events. The average expected background for that amount of data was $0.56^{+0.25}_{-0.13}$ events. CDF estimated the statistical significance of this excess to be 0.12. In other words, the probability of getting two or more dilepton candidate events on the assumption that only background processes are present is 0.12.

Two counting experiments searched for lepton plus jets events. The core of both was an algorithm to identify events with a *W* boson and three or more energetic jets. A *W* would be indicated by one energetic lepton and significant missing energy from an undetected neutrino. CDF found 52 such "*W* plus jets" events in the Ia data, where they expected approximately 46 from *W* production without top decays. To discriminate better against such background, they sought to single out events with a *b* quark, the other direct product of top decay. The two lepton plus jets searches used different means to "tag" events with *b* quarks.

Because the *W* boson has a large mass, much of the energy released in top decay would go into *W* production, leaving the accompanying *b* quark with little momentum. The *Soft Lepton Tagging (SLT)* counting experiment sought to capitalize on this by looking for events in which the *b* quark in turn decayed into a neutrino and a relatively low-momentum ("soft") lepton. The SLT analysis found seven candidate events, with an expected background of 3.1 ± 0.3 events. CDF estimated these results to have a statistical significance of 0.041.

The decay lifetime of the *b* quark is long compared to the lifetime of

4. Although I describe these search algorithms in loose, qualitative terms, they were defined by precise, complex criteria documented thoroughly in Abe et al. 1994.

the top quark itself. CDF had installed a very high resolution "silicon vertex detector" that could detect tracks of individual jet particles produced by *b* decay. The *Secondary Vertex (SVX) tagging* counting experiment sought events in which a jet originated from a point removed from the proton-antiproton interaction point (the primary vertex) by a distance consistent with the characteristic decay length of the *b*. The SVX search yielded 6 candidate events, with an expected background of 2.30 ± 0.29 events. The estimated statistical significance is 0.032.

CDF estimated the statistical significance of the combined results of the three counting experiments to be 2.6×10^{-3}.

Importantly for the analysis that follows, CDF developed three *different* SVX algorithms for tagging *b* quarks by looking for secondary vertices. The three algorithms, developed by distinct subgroups within the collaboration, were known as d–ϕ, jet probability, and jet vertexing. Only jet vertexing was used as a source of data for CDF's primary evidence claim yielding the statistical significance of 2.6×10^{-3}.

The three algorithms for SVX *b*-tagging were not entirely independent, as they all used data from the same apparatus. However, each algorithm used distinct methods for extracting a *b* tag from those data. Thus, each relied on certain assumptions not shared by the others. Next I describe some of the assumptions that each relied upon.

Jet vertexing tagged events by picking out tracks in the silicon vertex detector and requiring them to be fit to a secondary vertex significantly removed from the primary vertex. The algorithm required that the secondary vertex have $|L_{xy}|/\sigma_{L_{xy}} \geq 3.0$, where L_{xy} is the distance from primary to secondary vertex in the plane orthogonal to the beam line, and $\sigma_{L_{xy}}$ is the error on that quantity. Thus the use of jet vertexing to tag *b* quarks rests on the assumption that events lacking *b* quarks rarely yield results satisfying these constraints. Furthermore, the jet vertexing algorithm requires the sign of L_{xy} (determined by the sign of the dot product of the L_{xy} direction and the vector sum of the momenta of tracks in the tagged jet) to be positive. A further assumption is that among background events L_{xy} will be positive and negative approximately equiprobably.

The jet probability algorithm calculated for each track in a jet the impact parameter *d* (distance to the primary vertex at the nearest point extrapolated from the track). The algorithm then used that information to calculate a probability for each track in the jet, on the assumption that the track originated at the primary vertex. The probabilities for individual tracks were combined to form a joint probability for the jet as a whole. Jets with a very low joint probability on the assumption of having originated from the primary vertex were tagged. The probabilities used by this algorithm were drawn from a resolution function derived from the distribution of negative impact parameters in a sample of events called

the "50 GeV jet-trigger sample." Tagging with jet probability assumes the reliability of that distribution function, and hence that the sample on which it is based is unbiased.

The d–ϕ algorithm tagged events by means of the correlation between the impact parameter d and azimuthal angle ϕ in b decays. Here the relevant assumptions largely concern the geometry of secondary vertices from b decays. Specifically, "tracks with small d are likely to come from the primary vertex," whereas secondary vertices "will give tracks which form a line in the d–ϕ plane with non-zero slope" (Abe et al. 1994, 2988).

Although there is some overlap between the assumptions behind each method of b-tagging, each involves some assumptions that could fail without producing a failure of assumptions used in the other methods.

Now we are in a position to examine the use of robustness considerations in the analysis of the top quark evidence.

6.1. Convergence Amongst the b-Taggers: Second-Order Evidence. The distinction between first- and second-order evidence stands out prominently in CDF's discussion of the SVX counting experiment. CDF estimated the significance for the SVX search by itself to be 0.032. Do these results constitute at least some first-order evidence for the top quark hypothesis?[5] On the severe testing account employed here, the claim that they do constitute such evidence amounts to saying that the SVX results fit the hypothesis that there is a top quark, and that one would rarely get a result that fits as well if that hypothesis were false. But recall that evidential relationships themselves can be subjected to empirical inquiry on the present view. Hence one might seek evidence that the test to which the top quark hypothesis was subjected in the SVX counting experiment really was as severe as claimed by CDF, so as to secure that claim against being wrong in extent.

A successful argument from severity requires the elimination of the error of failing to satisfy the assumptions that underwrite the primary severity assessment. This can be done by testing those assumptions directly. CDF's "Evidence for Top Quark Production" paper is littered with such tests of their experimental assumptions, presenting a wide variety of second-order evidence, much of which does not rest on robustness considerations.

In complicated experimental endeavors investigators also worry about

5. As a matter of fact, the results of the SVX search alone would not have sufficed for CDF to go public with a claim to have evidence for the top quark. Even the combined statistical significance cited by CDF of 2.6×10^{-3} was thought by some to be insufficient for an evidence claim. Nonetheless, one can ask whether the results of one counting experiment constitute at least weak evidence for the top quark.

assumptions they have not thought to test directly. Skeptics might remain unpersuaded in the absence of evidence showing that the investigators who performed the experiment did not hastily cut off their testing of specific assumptions. In response, experimenters can provide evidence by showing that a similar result was obtained by a distinct procedure resting on independent assumptions.

For example, CDF built confidence in the SVX analysis based on jet vertexing by showing that the other secondary vertex *b*-tagging algorithms (the jet-probability and d–ϕ algorithms) lead to similar results. They noted that the jet-probability algorithm identified 4 events with a background of 2.3 ± 0.3, and that the d–ϕ algorithm tagged 5 events with an estimated 1.8 ± 0.2 background (Abe et al. 1994, 2994).

Furthermore, the three algorithms tagged some of the same events—in the actual data used for the top search, in data from control samples, and in samples of Monte Carlo–simulated data from top quark decays. CDF found that in the Monte Carlo–generated data, for example, "about 75% of the events tagged by the jet-vertexing algorithm are also tagged by at least one of the other algorithms, approximately 30% are tagged by both" (ibid.) They also noted that among the 6 actual candidate events in the real data identified by the jet-vertexing algorithm, 3 were tagged by the jet-probability algorithm, 4 were tagged by d–ϕ, and 2 were tagged by all three algorithms. They note that

> A study of the correlations among the different SVX tagging algorithms provides an additional check on whether the observed tags result from heavy-flavor [*b*-quark] jets or from the misidentification of light-quark or gluon jets. We have verified that there are large correlations among the algorithms for real heavy-flavor decays . . . . (Abe et al. 1994, 2995)

The agreement amongst algorithms thus provides evidence in support of a crucial assumption underlying their primary evidence claim: the reliability of jet vertexing as a procedure for *b*-tagging.

These correlations could not lend such support, however, if they would obtain no matter what. The agreement among the correlations becomes evidentially relevant only when it can be used to rule out certain kinds of errors. Here is where the requirement of discriminant validation has a role to play.

CDF employed discriminant validation in demonstrating that the algorithms failed to correlate in their "mistags," i.e., instances of tagging a secondary vertex that is not a result of a *b*-quark decay. Looking at events in a control sample, where they could measure the rate of mistaken *b*-tags, they found that events mistakenly tagged by jet vertexing were tagged by at least one of the other two algorithms about 20% of the time,

and by both of them only about 3% of the time (Abe et al. 1994, 2995). In other words, the results of the tagging algorithms tended *not* to agree when applied to events that did not contain *b* quarks.

The degree of convergence amongst results can itself be treated as evidence for an assumption about reliability: "If all six jet vertexing tags were due to tracking errors, we would expect approximately 1 of these events to be tagged by one of the other two algorithms. In contrast, five of the six events are tagged by at least one other algorithm" (Abe et al. 1994, 2995). CDF here argued that the agreement between the outcomes of the three algorithms (specified as the number of events tagged by two or more algorithms) was of a magnitude that far exceeded the expectation on the hypothesis that jet vertexing was tagging events incorrectly. In satisfaction of the severity requirement, the discrepancy between the observed degree of convergence and that expected if the assumption were false is of a magnitude that would be highly improbable if the assumption of reliable *b*-tagging were not true.

CDF's investigation of the relationships between the three secondary vertex algorithms exemplifies the distinction between first- and second-order evidence. The agreement between the outcomes of the three algorithms is presented, not as direct evidence for the top quark claim, but instead as evidence that jet vertexing tags the kind of events it is intended to identify. Conceivably, then (provided the other requisite assumptions are valid), it is evidence supporting the claim that the jet vertexing results are evidence for the top quark.

6.2. Combining Counting Experiments. The results of the three different *b*-tagging algorithms *could* have been combined into a single result supporting (as primary evidence) the top quark hypothesis. Doing so would have required a careful study of the considerable correlations between the three taggers, making a statistical assessment complicated. But such a single evidence claim based on the combined result is distinct from CDF's appeal to the robustness of the results from the three different algorithms as support for assumptions underlying the primary evidence claim.

CDF did combine the results of the three different counting experiments (SVX, SLT, and dilepton) to yield a significance estimate of 2.6×10^{-3}. Insofar as that significance estimate is intended as a partial indicator of the strength of the evidence for the top quark (strength of evidence being partly indicated by the low value of the significance), the fact that it combines results from somewhat independent tests cannot be appealed to as evidence for the claimed evidential strength of the combined result. The validity of the significance estimate for the combined result requires that each of the assumptions for each of the individual counting exper-

iments be valid. The fact that some of those assumptions are independent of one another does not help to show that this requirement has been met.

This is not to say that having a combined result from three different counting experiments was no better than having a result of comparable statistical significance from a single counting experiment. The SVX and SLT searches aimed at a decay channel distinct from that targeted by the dilepton search. Finding excesses in both decay modes helped to fix the theoretical interpretation of the statistical excess from the counting experiments in a way that was more probative than results from a single decay mode would have been. It is important, though, not to confuse the use of outcomes of different tests relating to different aspects of a complex phenomenon with the use of convergent results from tests employing independent assumptions as such. My focus here is on the latter issue.

CDF's use of the combined results from three counting experiments does illustrate the use of robust evidence to secure against categorical error. A statistical significance calculation based on just one testing procedure would be susceptible to some catastrophic failures that would not threaten a significance estimate based on combining results from several independent testing procedures. Failure of an independent assumption underlying the interpretation of one of those counting experiments would not entail the failure of assumptions underlying the other two. Hence, a weakened primary evidence claim might be sustainable in the face of the defeat of an assumption underlying the analysis of one of the counting experiments.

6.3. Second-Order Evidence or Back-Up Evidence? I mentioned in Section 3 that in some circumstances the same results can function either as second-order evidence or as first-order evidence. Deciding how to use convergent results may rest on deciding to use those results to address either of the two questions discussed in Section 4. This point can be seen clearly in another example of convergence in CDF's top quark data: the estimate of the top quark rest mass.

To determine the top mass, CDF examined events selected by either the SVX or SLT searches but also having a fourth jet meeting a relaxed energy threshold. Assuming that the 7 events that pass these criteria are in fact top quark events, they derive a mass estimate for the top quark for each event by reconstructing the kinematics of the event. CDF plots these estimates in two ways. The first consists of a simple histogram showing the numbers of events falling into 10 GeV/c^2 intervals of mass (see Figure 1), compared to similar distributions for Monte Carlo–generated background events (dotted line in Figure 1) and a combination of 175 GeV/c^2 top quark and background events (dashed line in Figure 1).

The second plot reflects the method used to arrive at a single best

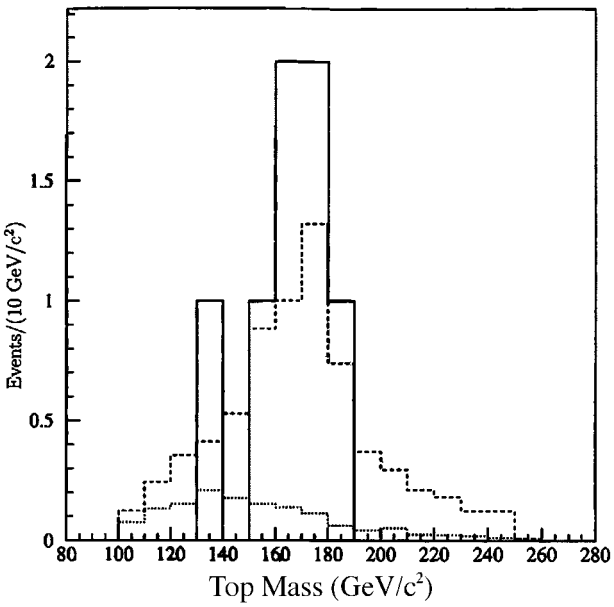

Figure 1. Distribution of mass estimates, comparing estimates for candidate events (solid histogram) and for a combination of Monte Carlo–generated background and top quark events (dashed histogram). The distribution for background alone is the dotted histogram (Abe et al. 1994, 3021).

estimate of the top mass based on the seven estimates derived from the individual events (see Figure 2). This curve shows the likelihood of a range of hypothesized top masses for the seven calculated top mass estimates. In other words, CDF sought to address the question: on the assumption of what hypothesis about the mass of the top quark is the probability of getting 7 events yielding these particular mass estimates maximized? They showed the probability of those results, for a range of mass hypotheses (equivalently, the likelihood of the hypotheses on the basis of those results) in terms of the negative logarithm of the likelihood, and then identified the mass hypothesis that minimized that quantity: $M_{\text{top}} = 174 \pm 10\text{GeV}/c^2$. (It should be noted with regard to discriminant validation that CDF also ran this mass reconstruction routine on Monte

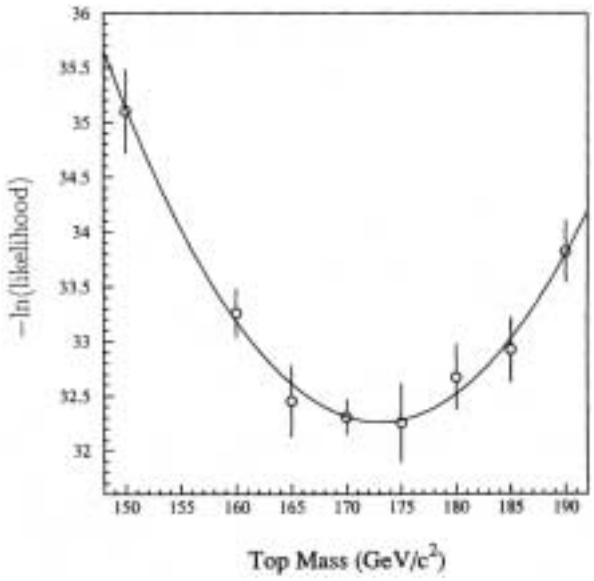

Figure 2. Negative log(likelihood) distribution for top mass estimates (Abe et al. 1994, 3020).

Carlo–generated background events passing their selection criteria; they found that these events yielded a "very broad peak centered at about 140 GeV/c^2" (Abe et al. 1994, 3019).)

Although these plots relating to the top mass estimate are clearly dependent on the SVX and SLT counting experiments, insofar as the estimates are based on events selected by those algorithms, they are independent in another sense. Indeed, it is their particular combination of dependence and independence in relation to the counting experiments that gives them their evidential resonance.

The events fed into the mass estimation procedure were all candidates selected by the SVX or SLT counting experiments. It is not, however, an assumption of the mass estimation procedure *as such* (as opposed to the procedure as a means of estimating the mass of some specified particle) that those counting experiments are not biased by having been tailored to increase the number of candidate events (a procedure known as "tuning on the signal"; see Staley 2002, 2004). As a procedure for estimating mass, the relevant assumptions concern the relationship between quantities like jet energy and the masses of hypothetical particles whose decay yields the quantities measured.

Biases in the counting experiments, however, were precisely what many

CDF members were worried about, and the peak in the mass estimate distribution was viewed as an independent test that was not susceptible to these same worries. An effort to tailor the cuts of the counting experiments to increase the number of candidate events would not in general be an effective way to produce a peak in the mass estimate distribution, if the events thus selected were mostly background. In a series of interviews I conducted in 1998, I asked a number of CDF physicists whether they believed that they really did have evidence for the top quark in their run Ia data, and quite a few of those I asked referred to the mass peak as an indication in favor of the evidence claim beyond the significance of the counting experiments alone. It is very improbable that background events would yield such a peak in either distribution.

Like the results from the other secondary vertex tagging algorithms, we can evaluate the results of the mass analysis with respect to their primary evidential status for the top quark claim. That is, we could ask whether the top quark claim passes a severe test with these particular results. Although they do not incorporate the result into their primary significance estimate, CDF notes that the likelihood-based estimate yields a result that "prefers the $t\bar{t}$ + background hypothesis over the background-only hypothesis by 2.3 standard deviations" (Abe et al. 1994, 3023). More importantly, by showing that the events selected by the SVX and SLT counting experiments were not just an oddball assortment of events with energetic jets that happened to be more numerous than expected, the mass peak helped convince some collaboration members that the cuts used in the counting experiments were not simply tailored to pick up additional candidate events, since the mass analysis was performed independently of choosing the counting experiment cuts. By supporting an important assumption behind the counting experiment significance calculation, the mass peak provided evidence that the counting experiment results were genuine, and not merely apparent, evidence for the top quark.

Most of the uses of convergent results I have discussed so far concern ways to address question (1) from Section 4. However, there is good evidence that CDF members were also concerned with question (2)—the more so as they worried about the adequacy with which type-(1) questions had been answered.

Some collaboration members had reservations about the evidence claims based on the counting experiments. A few had raised questions about potential biases from tuning on the signal. Some CDF physicists were satisfied that this was not a significant problem on the basis of considerations such as the mass peak. Others, who were not entirely satisfied by this argument, nonetheless accepted the soundness of the central claim of the "Evidence for Top Quark Production" paper, viz., that CDF had found evidence supporting the top quark hypothesis in their data.

Thus they agreed to the publication of the paper in spite of their reservations concerning some aspects of the analysis.

Of particular importance was the fact that the paper presented some kinds of evidence that were not included in the calculation of a statistical significance of the result based only on the counting experiments. Other features of the data that were generally thought to support the top quark hypothesis (such as the mass peak, certain kinematic features of the candidate events, or the fact that one dilepton candidate had a jet tagged by SVX and SLT) were officially treated as "checks" on the evidence from the counting experiments.[6]

Although such an appeal to checks indicates the use of these results as a source of second-order evidence for the top hypothesis, for some collaboration members skeptical about aspects of the counting experiments, these other features of the data served as back-up evidence. They reasoned that even if some parts of the counting experiments were biased, they could still truthfully claim evidence for the top quark because of the convergent results from other sources, the assumptions for which would remain undefeated by such bias. Thus for these physicists, the mass peak or the kinematic information functioned as an alternative source of primary evidence, rather than as second-order evidence. Rather than reassuring them about the "official" first-order evidence, it indicated to them that CDF's evidence claim would survive even if the assumptions behind the official version of it were not completely satisfied.

This suggests that robust evidence may have a special importance for collaborative experimental enterprises. The larger a collaboration is, the more likely it seems that group members with different perspectives will disagree over the validity of experimental assumptions. The pursuit of robust evidence thus may have a better chance of success than a "silver bullet" approach in those situations in which doubts arise over particular experimental assumptions. A multi-faceted pursuit may be less vulnerable to such doubts than a single-track approach, provided that the questions do not arise in every sector of the multi-faceted experiment.

This last point may seem to have only sociological or pragmatic, rather than epistemic, significance. For members of a large collaboration, this distinction is not so easily drawn. (I do not say that it cannot be drawn at all.) In a large collaboration such as CDF, the expertise required to evaluate an experimental result is widely distributed. This is why each

6. Some collaboration members, especially many of those who had worked on the kinematic analysis of the data, felt that the kinematic information should be included in the statistical significance calculation. Others felt that the systematic uncertainties were not sufficiently well understood for such inclusion. The dispute, in the words of a senior CDF member, "almost split the organization open" (Tollestrup 1995).

collaboration member has a say in the decision to publish a result. Any particular analysis may have been developed by a subgroup lacking expertise concerning some part of the detector of importance in producing relevant data, or in the application of techniques of data-refining on which their analysis depends.

Furthermore, different collaboration members will have different perspectives on the procedures by which an analysis was developed. The importance of this emerges clearly with regard to issues of potential bias. Group members who developed a set of cuts may, from first-hand knowledge, be in a position to assert confidently that those cuts were not chosen to include specific events as candidates. Colleagues not directly involved in those decisions may be justified in believing the cuts to be unbiased only insofar as they are justified in trusting those who chose the cuts.[7] Yet when a collaboration reaches the size of several hundred members, one cannot assume that each member knows every other member with much intimacy. There are various ways around this problem, such as the use of "blind" analysis techniques, but one way to strengthen one's justification in accepting the validity of a result is to make that validity insensitive to any particular assumption that can only be assigned a high degree of warrant by a portion of the group.

7. Conclusion. Advocates of robustness have maintained that robustness is evidentially relevant insofar as the reproducibility of a result by independent means indicates that the result is not an artifact of some particular process. My argument supports this assertion in one sense: robustness can enhance the security of an evidence claim.

The epistemic relevance of robustness can be seen clearly in CDF's argument, where it enhances the security of their evidence claim in several ways. (1) Robustness is appealed to as a means of providing second-order evidence, thus securing against an erroneous claim about the strength of their evidence. In order to demonstrate that the severe test requirement has been met here, CDF also uses discriminant validation. (2) The use of convergent results from counting experiments resting on distinct assumptions, when combined into a single result, also secures the first-order evidence claim against categorical failure due to defeated assumptions. (3) An erroneous claim about the strength of their evidence is avoided by having back-up evidence as security.

The error-statistical theory of evidence clarifies the constraints on the evidential value of robustness. Whether or not evidence is understood as error-statistical in nature, however, the uses of robust evidence described

7. See Krige 2001 for a discussion of the importance of trust in another particle physics collaboration experiment.

here make sense if evidence claims are understood as (1) claims about factual matters to be established empirically and (2) resting upon fallible assumptions about the procedures used to generate data and relate them to hypotheses of interest. I have shown here how robustness serves to make evidence claims more secure.

REFERENCES

Abe, F., M. G. Albrow, et al. [CDF] (1994), "Evidence for Top Quark Production in $\bar{p}p$ Collisions at $\sqrt{s} = 1.8$TeV", *Physical Review* D 50: 2966–3026.

Achinstein, Peter (1995), "Are Empirical Evidence Claims a Priori?", *British Journal for Philosophy of Science* 46: 447–473.

——— (2001), *The Book of Evidence*. New York: Oxford University Press.

Campbell, Donald T., and Donald W. Fiske (1959), "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix", *Psychological Bulletin* 56: 81–105.

Carnap, Rudolf (1962), *The Logical Foundations of Probability*, 2d ed. Chicago: University of Chicago Press.

Culp, Sylvia (1994), "Defending Robustness: The Bacterial Mesosome as a Test Case", in David Hull, Malcolm Forbes, and Richard Burian (eds.), *PSA 1994: Proceedings of the 1994 Biennial Meeting of the Philosophy of Science Association*. East Lansing, MI: Philosophy of Science Association, 46–57.

——— (1995), "Objectivity in Experimental Inquiry: Breaking Data-Technique Circles", *Philosophy of Science* 62: 430–450.

Franklin, Allan, and Colin Howson (1984),"Why Do Scientists Prefer to Vary Their Experiments?", *Studies in the History and Philosophy of Science* 15: 51–62.

Hacking, Ian (1983), *Representing and Intervening*. Cambridge: Cambridge University Press.

Hon, Giora (2003), "The Idols of Experiment: Transcending the 'Etc. List' ", in Hans Radder (ed.), *The Philosophy of Scientific Experimentation*. Pittsburgh: University of Pittsburgh Press, 174–197.

Hudson, Robert G. (1999), "Mesosomes: A Study in the Nature of Experimental Reasoning", *Philosophy of Science* 66: 289–309.

Krige, John (2001), "Distrust and Discovery: The Case of the Heavy Bosons at CERN", *Isis* 92: 517–540.

Mayo, Deborah (1996), *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Peirce, Charles S. (1992), *The Essential Peirce*, vol. 1. Edited by Nathan Houser and Christian Kloesel. Bloomington: Indiana University Press.

Staley, Kent W. (2002), "What Experiment Did We Just Do? Counterfactual Error Statistics and Uncertainties about the Reference Class", *Philosophy of Science* 69: 279–299.

——— (2004), *The Evidence for the Top Quark: Objectivity and Bias in Collaborative Experimentation*. New York: Cambridge University Press.

Tollestrup, Alvin (1995), interview by the author. Tape recording. 11 October, Fermilab, Batavia, IL.

Trout, J. D. (1993), "Robustness and Integrative Survival in Significance Testing: The World's Contribution to Rationality", *British Journal for the Philosophy of Science* 44: 1–15.

Whewell, William (1989), *Theory of Scientific Method*. Edited by Robert E. Butts. Indianapolis: Hackett Publishing Company.

Wimsatt, William (1981), "Robustness, Reliability, and Overdetermination", in Marilynn B. Brewer and Barry E. Collins (eds.), *Scientific Inquiry and the Social Sciences*. San Francisco: Jossey-Bass, 124–163.