NATIONAL INTELLIGENCE UNIVERSITY

(U) What This Is:

(U) *Research Shorts* fuse two research cultures, blending intelligence information with academic insights on topics of interest to the IC. *Shorts* are intended to constructively start and add to the IC's conversations—not to finish them. NIU, the sole fully TS/SCI-cleared university, publishes the *Shorts*.

(U) What This Is Not:

(U) *Research Shorts* are not finished intelligence and are not IC-coordinated. The opinions expressed are solely those of the author and do not represent those of any U.S. Government agency.

RESEARCH SHORT

CATALYST Sparking Constructive Conversations About the Future of Intelligence

May 16, 2019

IMAGE FROM SHUTTERSTOCK

(U) Should Confidence Levels Be Reinforced, Refined, or Replaced?

(b)(6)
[redacted]

(U//~~FOUO~~) The IC has long attempted to accurately characterize and convey analytic uncertainty for its national security customers, but decades of debate, reworked guidance, and trial-and-error application have left confidence levels ill-defined, misunderstood, and often misapplied. Please participate in the authors' survey on refining or replacing confidence levels. Should they be used? If so, how? And, if not, what might be used in their stead? The authors plan to use your input to produce a second *Research Short* assessing the survey results and insights offered.

(U) Time for a Change

(U//~~FOUO~~) It is time to give serious thought to refining or replacing confidence levels. In theory, confidence levels allow analysts to produce assessments that acknowledge levels of uncertainty in standard terms that all consumers can understand. Explicit criteria for high, moderate, and low confidence levels enable analysts to measure and consistently convey a degree of certainty in data, sources, assumptions, and methodology used to support analytic judgments. This use of confidence levels draws on well-defined practices in the social sciences and permits differentiation of the certainty associated with the support for dissimilar judgments.

(U) In application, however, the IC generally has not used confidence levels for their original and intended purpose, and corrective efforts have largely failed to resolve the most significant challenges associated with their use. For one, the distinction between the basis of an assessment of an outcome's probability and the confidence levels that convey uncertainty about that assessment is unclear because the factors important for one measure are equally important for the other (see box). This confuses consumers and analysts as to what a confidence level conveys, particularly as IC organizations differ in their definition and use of confidence levels. Critics also charge that confidence levels convey a false sense of precision and misapply a tool appropriate only for statistics.[1] In addition, the new threat environment, as well as the possibilities offered by artificial intelligence and advanced analytics to calculate and convey uncertainty, demands a fresh look at this longstanding challenge.

(U) PROBABILITY JUDGMENTS AND CONFIDENCE LEVELS: SEPARATE BUT NOT SEPARABLE

(U) Paired judgments about the probability of a future event and the level of confidence in that probability estimate are distinct but related, as demonstrated by the following hypothetical examples. The first is plausible, the second nonsensical:

- (U) Sylvania probably will invade Freedonia tomorrow. We have moderate confidence in this judgment, because . . .
- (U) Sylvania almost certainly will invade Freedonia tomorrow. We have low confidence in this judgment, because . . . [2]

(U) This explanation is drawn from an earlier *Short*, "**(U) Improving IC Analytic Confidence Levels**"—one of several articles in an ongoing exploration of intelligence tradecraft in the NIU *Research Short* series.

(U) Ultimately, deciding whether the IC continues to revise and reinforce its guidance on the use of confidence levels, refines their basis, or replaces them with another construct to capture and convey uncertainty is likely to be a step in an ongoing discussion. The evolution of the IC's use of confidence levels over time illuminates the many difficulties inherent in their application and the challenges to providing proper guidance on their implementation. To advance the discourse, the authors offer several options for moving forward and ask readers to offer their thoughts in a short survey that will inform further research.

(U) Muddled Application Spawned Confusion

(U) The inconsistent application of confidence statements and accompanying implementation guidance—often rolled out only after problems had occurred—calls attention to the confusion over their use. A review of intelligence analysis reveals only limited use of explicit confidence levels (high, moderate/medium, or low) before the controversy surrounding a 2002 study on WMD in Iraq.

(U) Confidence levels during the 1950s and 1960s conveyed margins of error in estimated numbers or were applied in general terms to estimative judgments through terms such as "possible" and "probable." During the next three decades, the IC only sporadically incorporated confidence assessments focused primarily on sourcing material or key judgments. IC estimative judgments in the early 2000s increasingly stated confidence levels, but without clear definition or criteria. Among the National Intelligence Estimates (NIEs) that contained unexplained confidence language was the October 2002 NIE *Iraq's Continuing Programs for Weapons of Mass Destruction*—the flawed analysis preceding the 2003 Iraq War and fruitless search for WMD.[3] (For more information on the history of confidence statements, see Appendix A.)

Implementing ICD 203's guidance has been a struggle because the ICD did not define confidence levels.

(U) The 9/11 attacks and WMD NIE prompted Congress to mandate in the 2004 Intelligence Reform and Terrorism Prevention Act (IRTPA) that IC products be evaluated for their analytic tradecraft, including whether they "properly caveat and express uncertainties or confidence in analytic judgments."[4] The 2007 Intelligence Community Directive (ICD) 203 *Analytic Standards* codified IRTPA's mandate, directing that IC analytic products indicate "the level of confidence in analytic judgments and explain the basis for ascribing it," flagging any "sources of uncertainty."[5]

(U) Implementing ICD 203's guidance has been a struggle because the ICD did not define confidence levels, leaving analysts and agencies to inconsistently delineate the term; the criteria upon which the confidence levels should be based; the difference between confidence levels and likelihood estimative language; and guidelines for when and how to incorporate confidence levels into IC products. Although most organizations agreed on the three confidence levels (high, moderate or medium, and low) established by the National Intelligence Council (NIC), they disagreed on the criteria for determining the levels. All placed the primary emphasis on the quality of the sourcing, but some also weighed the number and nature of assumptions made; the difficulty of the question being addressed; and the potential for deception.[6] (For more detail on the IC's guidance on the use of confidence judgments, please see Appendix B.)

(U//~~FOUO~~) IC policy customers also have labored to understand what confidence levels were conveying. Questions from senior consumers highlighted confusion when intelligence assessments put likelihood language and confidence in the same sentence.[7]

(U) Reinforce, Refine, Replace, or Do Nothing?

(U//~~FOUO~~) During the past 15 years, the more frequent use of confidence level language to convey uncertainty in IC analytic assessments and in the sourcing underlying them has created multiple problems for analysts and their customers. Numerous revisions of the guidance on how to convey confidence levels have failed to unequivocally define confidence levels and their basis or demonstrate that they provide accurate insight into the uncertainty surrounding judgments and assessments. These years of debate, revised guidance, and trial-and-error application have left the IC with more questions than answers on the use of confidence levels and the deeper concept of

uncertainty. Entangled in these questions are issues fundamental to modernizing intelligence, such as the tradecraft balance between analysis and production and the skills the IC workforce needs to handle uncertainty appropriately. Moving forward requires the IC to answer simple yet potentially game-changing questions—should the community use confidence levels at all? Is this the best or even an adequate construct for maintaining rigor in intelligence analysis?

(U//~~FOUO~~) More study on the utility of confidence levels is needed to answer those questions. In the meantime, the IC should consider scenarios for ending the near continuous discord on the question of confidence levels. Although general in nature, discussion of four options for the IC—to **reinforce** the use of confidence levels by improving guidance; to **refine** confidence levels to improve their rigor; to **replace** them with another construct; or to **do nothing**—could identify and clarify some fundamental points of dissent regarding their use.

(U) Reinforce: Keep the Construct, Improve the Guidance

(U//~~FOUO~~) In this scenario, confidence levels are broadly accepted as an appropriate construct for expressing uncertainty, and current tradecraft guidance—ICD 203 and individual agency directives—is revised to clearly define and synchronize the rigorous use of confidence levels across the community. Clarifying criteria for the different levels or techniques might minimize what appears to be a default to "moderate" for most products. Broad community buy-in would be needed to keep this from being just another iteration of the update/rework cycle. Therefore, collaborative development and testing, with the broadest level of participation, are essential—particularly by IC methodologists and tradecraft specialists, whose perspective will be crucial to convincing leadership and colleagues to follow the revised tradecraft. Doing a "beta" test that allows for the possibility of failure and nonadoption before any guidance is finalized would build the credibility needed to prevent yet another cycle of refuting and rejection.

(U//~~FOUO~~) This scenario's linchpin assumption is that no serious overhaul of the conceptual basis of the confidence level construct is needed because the theoretical framework behind how it assesses and expresses uncertainty is considered sound. Indications suggest, however, that this is exactly where the problem with confidence levels lies.

(U) Refine: Keep the Construct, Improve the Rigor

(U//~~FOUO~~) The "refine" scenario would arise from a successful challenge to the conceptual framework of the construct itself that called into question the rigor of any finding it generates. Examples of this kind of challenge include determining that the definition of confidence fails to capture a critical aspect of certainty or arguing that the levels cannot effectively indicate significant changes in certainty. In this case, the construct must be deconstructed and rebuilt to correct the flaw. The basis for this questioning might stem from the idea that the construct's quantitative roots in statistics and modeling have been wrongly applied to the IC's more qualitative analysis. The original tool's quantitative formulas and reasoning supporting the levels and overall assessment get lost in the change to qualitative adaptation. That does not automatically equate to a lack of rigor, but it does require a reclarification of how the tool now leads from information input to

a judgment call. For some in the community this has not happened, and the result is distrust in the methodology behind the tool.

(U//~~FOUO~~) The linchpin assessment here is that the community will elect to maintain the construct as the best option for handling uncertainty.

(U) Replace: Adopt a Different Means To Convey Uncertainty

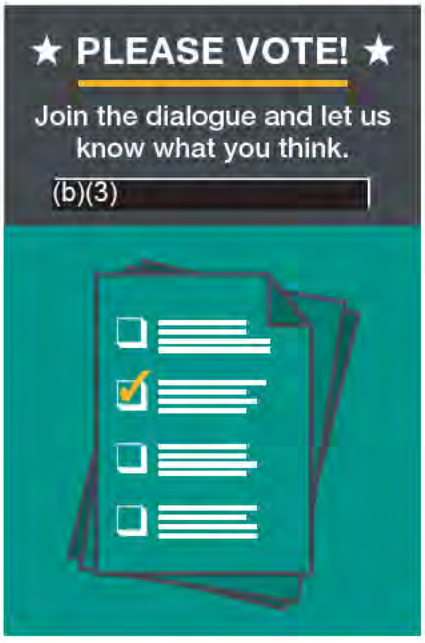
(U//~~FOUO~~) This is a more feasible scenario than it might appear at first. Confidence level is removed from the list of approved applications for meeting the analytic tradecraft standard that requires analysts to explain the impact of the sources of uncertainty behind an assessment and is replaced by any of a number of approaches already being explored in the community. Some possibilities for replacing confidence levels include determining that estimative language without confidence level is sufficient or choosing to use a percentage-based approach to express uncertainty.

(U//~~FOUO~~) Estimative language—the use of words such as "probably" or "likely" to express the likelihood of an outcome—has been an almost constant companion of confidence levels in the IC. Sometimes the partnership has been overly tight; some earlier applications conflated the concepts to the point of creating unnecessary correlations in language—for example, requiring "low" confidence for an assessment of "will not occur." Recently that partnership has become noticeably uneasy as some practitioners question the need for both constructs to adequately express uncertainty. Some of these analysts also advocate for Bayesian or other probability theory-based approaches to replace confidence levels and qualitative methods for assessing uncertainty in general. This proposal brings up not just questions about the feasibility of meaningful quantitative assessment against persistent or enduring uncertainty, but also the potential need to reskill the current analytic workforce to accurately apply such methods.

(U) What Next?

(U//~~FOUO~~) Regardless of the eventual outcome, the governing drivers of the future of confidence levels or any other approach to assessing uncertainty will need to include buy-in from the community, the level of analytic rigor perceived or assessed in application, and ease in adoption across the IC workforce.

(U//~~FOUO~~) Please join the dialogue by taking a short online survey through this link: (b)(3) Do you agree with one IC analyst who remarked in a 2012 (b)(3) comment: "When we apply techniques which are at some level inadequate to the task, we sometimes get answers which don't make much sense." Or is your thinking more in line with another analyst who responded in the same blog, "I am not sure that we can devise a doctrine for the expression of confidence that will elegantly deal with the problem in a universally satisfactory way…



Perhaps it would be better to focus on trying to identify those domains on which we can usefully express degrees of confidence and those we can't."[8]

(U//~~FOUO~~) Please indicate whether and why the IC should:

1. Reinforce current confidence level guidance and require all IC organizations to employ confidence levels.
2. Refine the confidence level construct.
3. Replace confidence levels with another construct to measure and convey uncertainty.
4. Do nothing at all.

(U//~~FOUO~~) The authors plan to use your input to produce a second *Research Short* assessing the results and insights offered.

(b)(6)
[redacted]

ICOD: April 30, 2019

(U) If you have comments, questions, or a suggestion for a Research Short topic or article, please contact the NIU Office of Research at (b)(3) [redacted]

(U) Appendix A

(U) Confidence Levels Evolved Slowly

(U) The historic irregularity of explicit confidence levels in IC products may reflect inattention from the "father of intelligence analysis," Sherman Kent, who included no discussion of "confidence" language in his seminal work, "Words of Estimative Probability," published in *Studies in Intelligence* in 1964.[9] Nor is "confidence" found in IC glossaries of intelligence terms and definitions compiled in the late 1970s and 1980s.[10]

(U) Intelligence analysis originally applied confidence levels to estimated numbers in accordance with statistical teachings on margins of error. "Confidence in NIE-65 Production Statistics," a 1953 paper supporting NIE 65 *Soviet Bloc Capabilities Through 1957*, noted, for example, that "a more reliable" picture of Soviet economic production statistics could be achieved by "recasting each estimate to reflect the relative degree of confidence in that particular figure."[11] Generic "confidence" statements, flagged through words including "possible" or "probable," were cited in a 1966 *Studies in Intelligence* article on military-economic analysis in NIEs.[12] This practice may have reflected the thinking of another early intelligence practitioner, BG Washington Platt, whose Principle of Degree of Certainty required consideration of the reliability of statements of fact, the precision of quantitative data, and the probabilities of estimates and conclusions. As Platt said,

> (U) In each case the degree of reliability, or of precision, or of probability may be very high or very low or in between. According to this principle, one of the essential responsibilities of the intelligence production officer is to determine by critical study the reliability, precision or probability, as the case may be, of each important element of his paper and then to make this clear to the reader. This procedure greatly increases the usefulness of any paper.[13]

(U) DIA's experimental effort in 1976 to use percentages to convey the probability that a given judgment was valid and letters to convey the analyst's confidence in the source material was an example of the IC's sporadic attempts to express analytic uncertainty with "confidence."[14] Even after DIA discontinued the pilot, subsequent products did note that confidence in the supporting evidence was taken into account in making estimative judgments in numeric form (e.g., chances out of 10 or percentages).[15]

(U) Other IC entities also incorporated confidence in assessing the uncertainty involving key judgments. When confronted with a contentious NIE on Soviet offensive chemical weapons in October 1984,[16] the Weapons and Space Systems Intelligence Committee produced a "Confidence Level Table" that depicted commensurate likelihood expressions with numeric probability ranges.[17] Over the next 15 years, "confidence" continued to be used generically in conjunction with estimative judgments. A 1995 CIA tradecraft note was indicative of this approach.[18] The explicit confidence levels associated with key estimative judgments included in IC products between 2000 and 2002—including the October 2002 WMD NIE—remained ill-defined.[19]

(U) Appendix B

(U) Official Guidance on Confidence Levels Often Reactionary and Ill-Defined

(U) The NIC's "What We Mean When We Say" textbox was among the first IC efforts to implement the 2004 IRTPA mandate that IC products "properly caveat and express uncertainties or confidence in analytic judgments."[20] The textbox—included in all NIC products—identifies the estimative and confidence level language used in each product and the criteria for each term. Agency-specific "Tradecraft Notes" were drafted to help IC analysts and analytic managers properly apply the mandate in the 2007 ICD 203 *Analytic Standards* to explain the basis for ascribing confidence in a product. ICD 203 specified that "sources of uncertainty—including information gaps and significant contrary reporting—should be noted and linked logically and consistently to confidence levels in judgments" and, as appropriate, products "should identify indicators that would enhance or reduce confidence or prompt revision of existing judgments."[21]

(U//~~FOUO~~) The NIC textbox, ICD 203, and the many tradecraft notes failed to dispel disagreements over whether or how estimative likelihood language differed from confidence levels. Multiple (b)(3) blogs between 2008 and 2014 document divergent views within the IC. A contributor to a 2012 blog addressing "What Would You Fix or What is Missing in the IC Analytic Standards" recommended that ICD 203 should "at a minimum differentiate between confidence and probability with more precise wording and definitions."[22] Another analyst opined "part of the confusion may be we are trying to communicate on one scale concepts that require at least two scales."[23]

(U//~~FOUO~~) Frustration expressed by senior consumers about confusing confidence statements in IC assessments prompted Robert Cardillo, then Deputy Director of National Intelligence for Intelligence Integration (DD//II), to issue guidance in 2013 on the purpose of statements of probability and analytic confidence and how they differed. He argued IC customers are best served when analysis distinguishes between statements of probability—or the likelihood that an event will occur—and analytic confidence—the degree of conviction in an analytic judgment, primarily reflecting the level of trust placed in the sources. To emphasize the duality, products prepared by the NIC and material included in the President's Daily Brief were directed to present probability and associated confidence levels in separate statements, although other "authoring elements" retained the discretion whether to use confidence levels at all. Cardillo promised that separating statements of probability from confidence levels would become the IC standard under an updated ICD 203.[24]

(U//~~FOUO~~) Although the DD/II's guidance appeared in the January 2015 ICD 203 as part of its treatment of the second analytic tradecraft standard addressing uncertainty, the changes did not curtail the confusion or discord. The NIC asked a contractor to examine the use of confidence levels to help clarify what confidence levels represented and how they should be used.[25] The IC's Analytic Standards Evaluation Action Group (ASEAG) also asked a sub-working group to identify better methods to convey uncertainty.[26] A plethora of revised or new tradecraft notes from the NIC, DIA, the Department of Energy, and others—as recent as December 2018—strongly suggest the IC remains confused over how and when to use confidence levels to capture and express uncertainty.[27] [28] [29] [30]

(U) Endnotes

1 (U) Kristan J. Wheaton, "The Revolution Begins on Page Five: The Changing Nature of NIEs," *International Journal of Intelligence and Counterintelligence* 25, no. 2, 339–41.

2 (U) NIU; Research Short by (b)(6) ; 11 DEC 2018; (U) Improving IC Analytic Confidence Levels; Classification of extracted information is U//FOUO; Overall classification is U//FOUO; (b)(3)

3 (U) National Intelligence Council, "Iraq's Continuing Programs for Weapons of Mass Destruction," October 2002, FOIA Collection, Freedom of Information Act Electronic Reading Room, 0001075566.

4 (U) Intelligence Reform and Terrorism Prevention Act of 2004, Pub. L. No. 108-458, 118 Stat. 3638 (2004), Section 1019, https://www.gpo.gov/fdsys/pkg/PLAW-108publ458/pdf/PLAW-108publ458.pdf.

5 (U) Office of the Director of National Intelligence, *Intelligence Community Directive 203: Analytic Standards*, June 21, 2007, available online in the Homeland Security Digital Library at https://www.hsdl.org/?view&did=479263.

6 (U) DIA; DI Tradecraft Note 1-08; JUL 2008; (U) Calibrating Analytic Confidence; Classification of extracted information is U; Overall classification is U; (b)(3)

7 (U) ODNI; AIS-AA-2014-05; OCT 2014; (U) Ukraine—Russia Crisis: Consumer Feedback on Quality of Intelligence Community Analysis and Support; p. 6; Classification of extracted information is U; Overall classification is S//NF. See consumer comments for an example; (b)(3)

8 (b)((3), (b)(6)

9 (U) Sherman Kent, "Words of Estimative Probability," *Studies in Intelligence* 8, no. 4 (Fall 1964), 49-65, https://www.cia.gov/library/center-for-the-study-of-intelligence/kent-csi/vol8no4/pdf/v08i4a06p.pdf.

10 (U) Director of Central Intelligence, Intelligence Community Staff, *Glossary of Intelligence Terms and Definitions*, June 15, 1978, General CIA Records Collection, Freedom of Information Act Electronic Reading Room, CIA-RDP80M00596A000500020003-7.

11 (U) Memorandum for Assistant Deputy Director/Intelligence and Assistant Director, National Estimates, "Confidence in NIE-65 Production Statistics," April 8, 1953, General CIA Records Collection, CIA Freedom of Information Act Electronic Reading Room, CIA-RDP79-01206A000200010010-9.

12 (U) Julie O. Kerlin, "Military-Economic Estimating: A Positive View," *Studies in Intelligence* 10 (Fall 1966), 35-44, CIA Freedom of Information Act Electronic Reading Room, FOIA Collection 0000609095, DOC_0000609095.pdf.

13 (U) Washington Platt, *Strategic Intelligence Production: Basic Principles* (New York: Frederick A. Praeger, 1957), 44.

14 (U) Director of Central Intelligence, Intelligence Community Staff, *Review of National Intelligence*, August 1976, General CIA Records Collection, CIA Freedom of Information Act Electronic Reading Room, CIA-RDP86B00269R001200200001-5.

15 (U) Defense Intelligence Agency, Defense Intelligence Estimate DIE SOV 2-76, "Military Significance of Soviet Developed Facilities in Somalia," February 20, 1976, available online in the Foreign Relations of the United States Series, 1969-1976, Volume E-6, Documents on Africa, 1973-1976, Document 155, http://history.state.gov/historicaldocuments/frus1969-76ve06/d155.

16 (U) Director of Central Intelligence, "Special National Intelligence Estimate: The Soviet Offensive Chemical Warfare Threat to NATO," October 1984, CIA General Records Collection, CIA Freedom of Information Act Electronic Reading Room, CIA-RDP87T00217R000200070009-1.

17 (U) National Foreign Intelligence Board, Memorandum for National Foreign Intelligence Board Principals, "Weapon and Space Systems Intelligence Committee Confidence Level Table," November 8, 1984, CIA General Records Collection, CIA Freedom of Information Act Electronic Reading Room, CIA-RDP86M00886R000200150019-7.

18 (U) CIA Directorate of Intelligence (b)(3) ; MAR 1996; (U) A Compendium of Analytic Tradecraft Notes, Volume 1 (Notes 1-10); p. ix, 7, 15, 38-9; Classification of extracted information is U; Overall classification is U; (b)(3)

19 (U) National Intelligence Council, "Iraq's Continuing Programs for Weapons of Mass Destruction," October 2002, FOIA Collection, Freedom of Information Act Electronic Reading Room, 0001075566.

20 (U) Intelligence Reform and Terrorism Prevention Act of 2004, Pub. L. No. 108-458, 118 Stat. 3638 (2004), Section 1019, https://www.gpo.gov/fdsys/pkg/PLAW-108publ458/pdf/PLAW-108publ458.pdf.

21 (U) Office of the Director of National Intelligence, *Intelligence Community Directive 203: Analytic Standards*, June 21, 2007, available online in the Homeland Security Digital Library at https://www.hsdl.org/?view&did=479263.

22 (U)(b)(3) ; Analytic Methods Forum Discussion; 2012–13; (U) What Would You Fix or What is Missing in the IC Analytic Standards?; Classification of extracted information is U;(b)(3)

23 (U)(b)(3) Analytic Methods Forum Discussion; 2012–13; (U) What Would You Fix or What is Missing in the IC Analytic Standards?; Classification of extracted information is U;(b)(3)

24 (U) ODNI; Deputy Director of National Intelligence for Intelligence Integration Memorandum E/S 00387; 13 JUN 2013; (U) New Guidance on Statements of Probability and Confidence; Classification of extracted information is U//~~FOUO~~, Overall classification is U//FOUO;(b)(3) .

25 (U) Defense Group Inc., "Confidence in Intelligence, Status Summary," January 2017, Briefing Slides.

26 (U) ASEAG, ATS-2 Working Group; Briefing; 8 NOV 2016; (U) Convey Analytic Uncertainty; Classification of extracted information is U; Overall classification is U//FOUO.

27 (U) NIC; TGM 2018-001; 15 JUN 2018, (U) Tradecraft Memorandum: Using Confidence Levels in NIC Products; Classification of extracted information is U; Overall classification is U//FOUO;(b)(3)

28 (U) DIA; Tradecraft Note 04-18; 21 SEP 2018; (U) Uncertainty: Expressing and Explaining Analytic Uncertainty in Defense Intelligence Analysis; Classification of extracted information is U; Overall classification is U;(b)(3)

29 (b)(3)

30 (U) NIU; Research Short by(b)(6) 11 DEC 2018; (U) Improving IC Analytic Confidence Levels; Classification of extracted information is U//FOUO; Overall classification is U//FOUO;(b)(3)