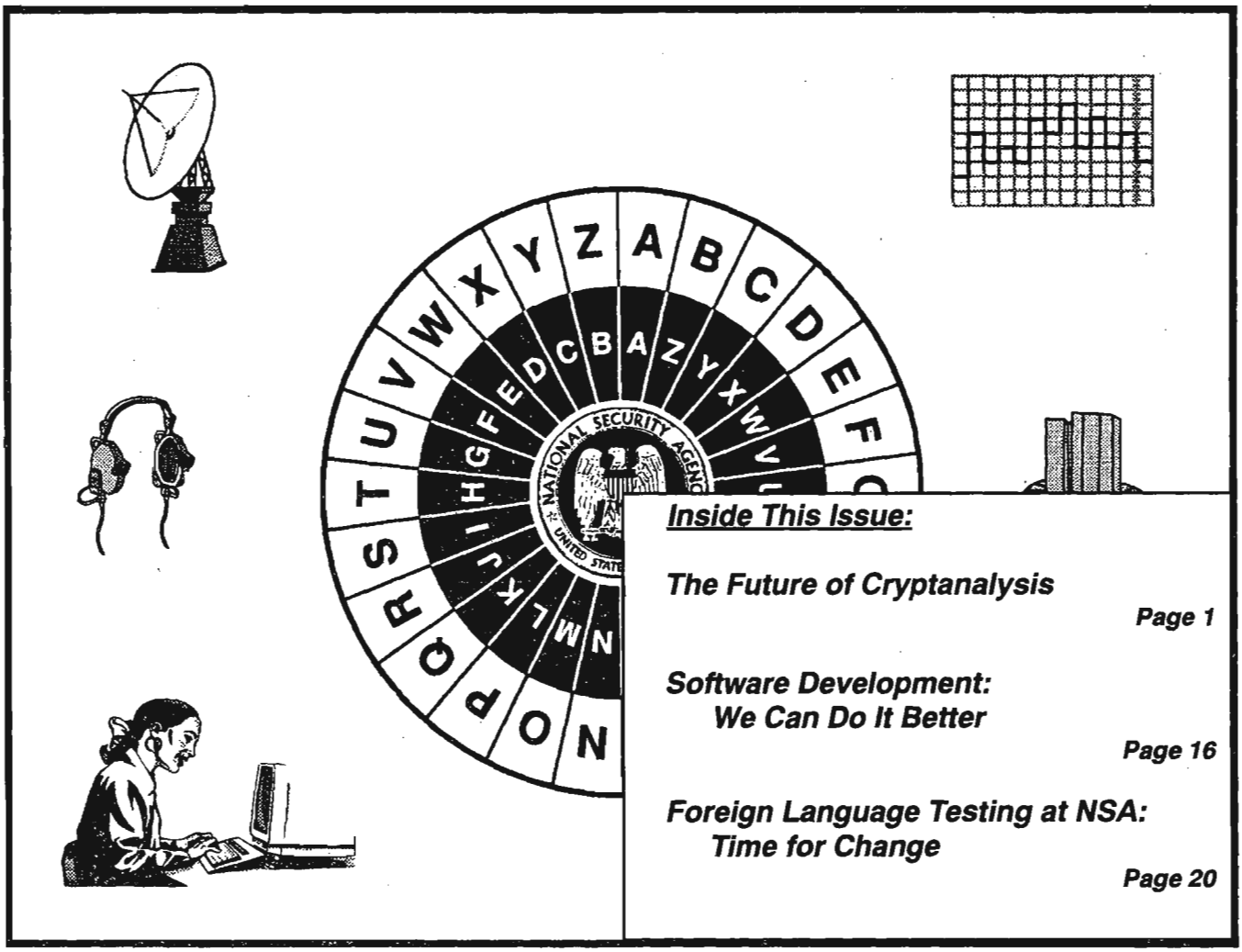
~~TOP SECRET~~

NATIONAL SECURITY AGENCY

CRYPTOLOG

The Journal of Technical Health

Vol. XX, No. 3 FALL 1995



Inside This Issue:

~~Derived From: NSA/CSSM 123-2 Dated 3 September 1991~~
~~Declassify On: Source Marked "OADR" Date of source: 3 Sep 91~~

~~THIS DOCUMENT CONTAINS CODEWORD MATERIAL~~
~~TOP SECRET~~

CRYPTOLOG

Fall 1995
Vol. XX, No. 3

Published by P05, Operations Directorate Intelligence Staff

To submit articles and letters, please see last page

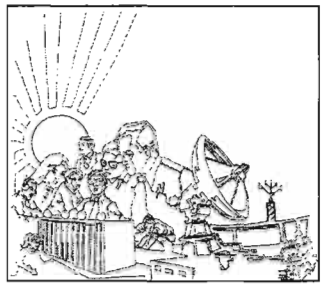


Table of Contents

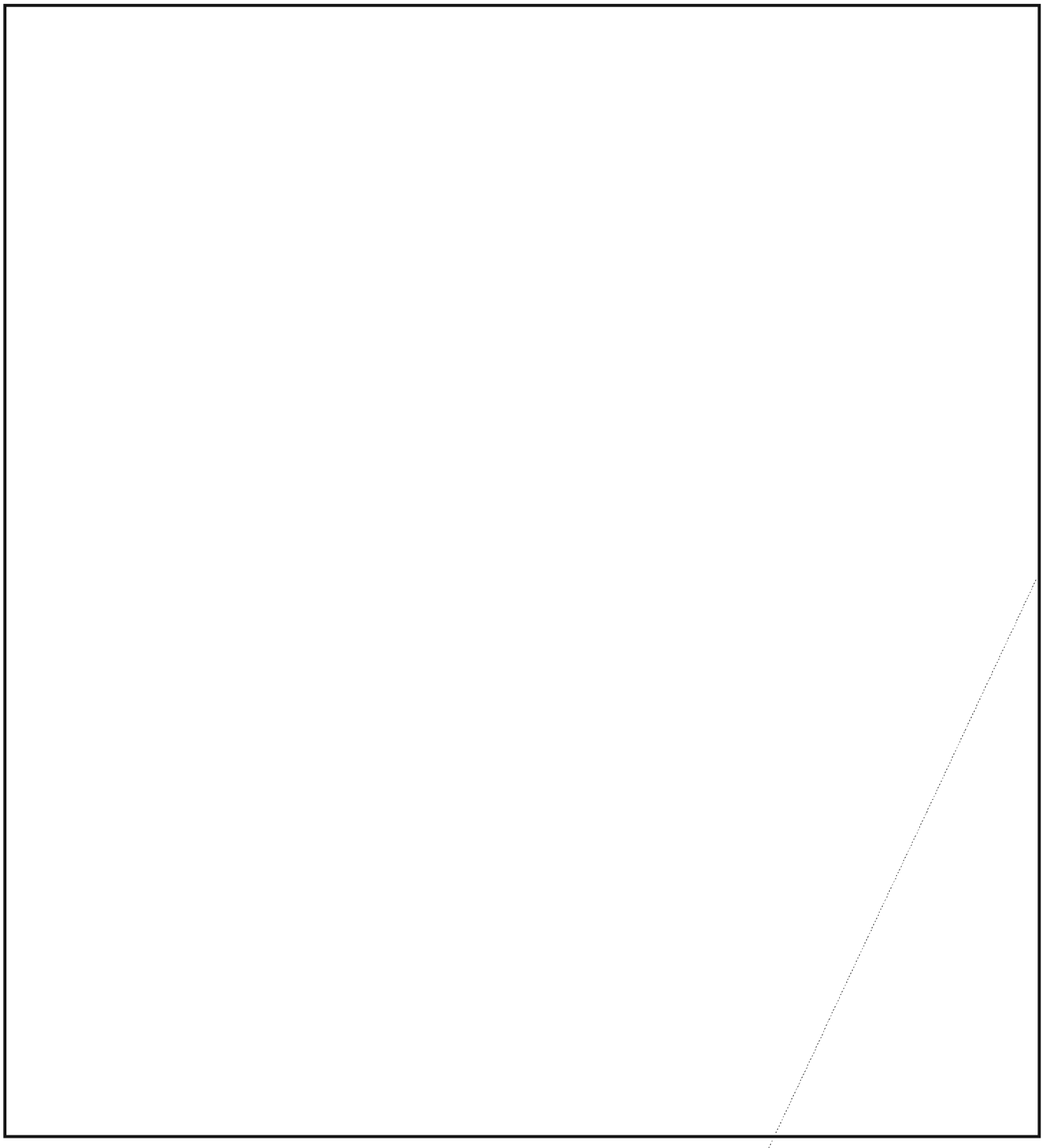
P.L. 86-36

EO 1.4.(c)
P.L. 86-36

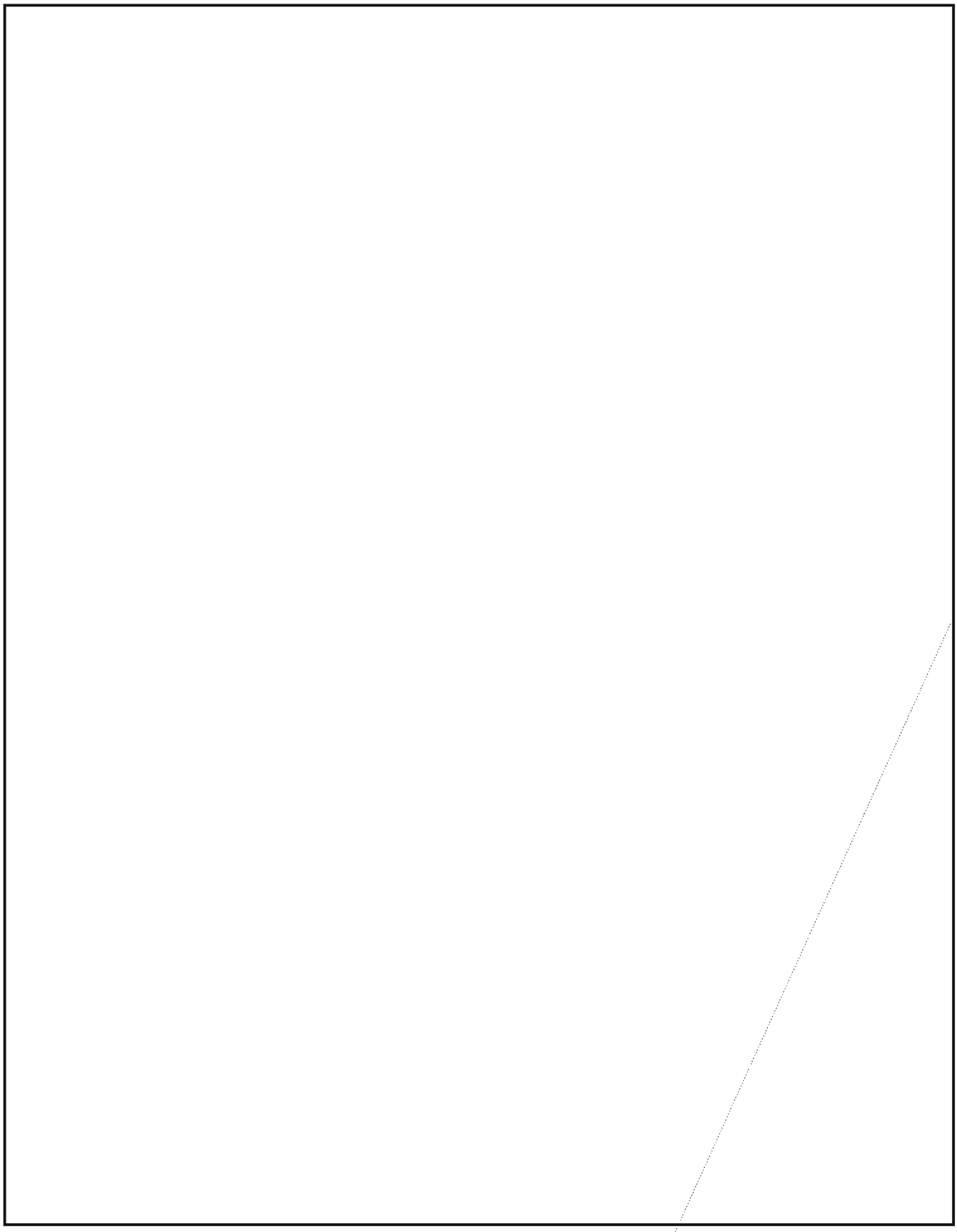
The Future of Cryptanalysis

by [redacted] Chief Z

P.L. 86-36



EO 1.4.(c)
EO 1.4.(d)
P.L. 86-36



EO 1.4.(c)
EO 1.4.(d)
P.L. 86-36

Tech Trend Notes publishes a Calendar of Events sponsored by NSA, academia, and professional associations. Here's a sample of what's happening:

Event	Date	Location	Where to call:
Defense Information Warfare Symposium	11-12 Dec	New Orleans, LA	(703) 681-1346
4th Intl. World Wide Web Conference	11-14 Dec		Web site: URL: http://www.crs.lcs.mit.edu/registration-form.html
RSA Data Security Conference	17-19 Jan	San Francisco, CA	(415) 595-8782
DoD Photonics Conference	26-28 Mar	McLean, VA	(703) 631-6128
TechNet Canada '96: "Gov't. and Industry Info Exchange"	17-18 Apr	Ottawa, Canada	(613) 563-0093
Eurocrypt '96	12-16 May	Zaragoza, Spain	email: sec96@aegean.ariadne-t.gr

Reinvention Laboratory for SIGINT Reporting

NO MORE DEFAULTING
TO THE OLD, ENTRENCHED WAYS OF DOING BUSINESS
(" . . . the way we've always done it . . . ")

Officially designated as a reinvention lab (a product of the "reinventing government" initiative), P054 has been given the go-ahead to undertake experiments in reporting. In a continuing search to identify more effective ways of getting SIGINT to customers, P054 will use the ***SIGINT Digest*** and NSA Broadcast Network as test-beds for trying out new reporting styles, content, and dissemination methods, with potential future application by Agency reporting elements.

Your suggestions are welcome.
Please contact [redacted] P.L. 86-36
acting Research Director, at 963-3123s.

MACHINE TRANSLATION:

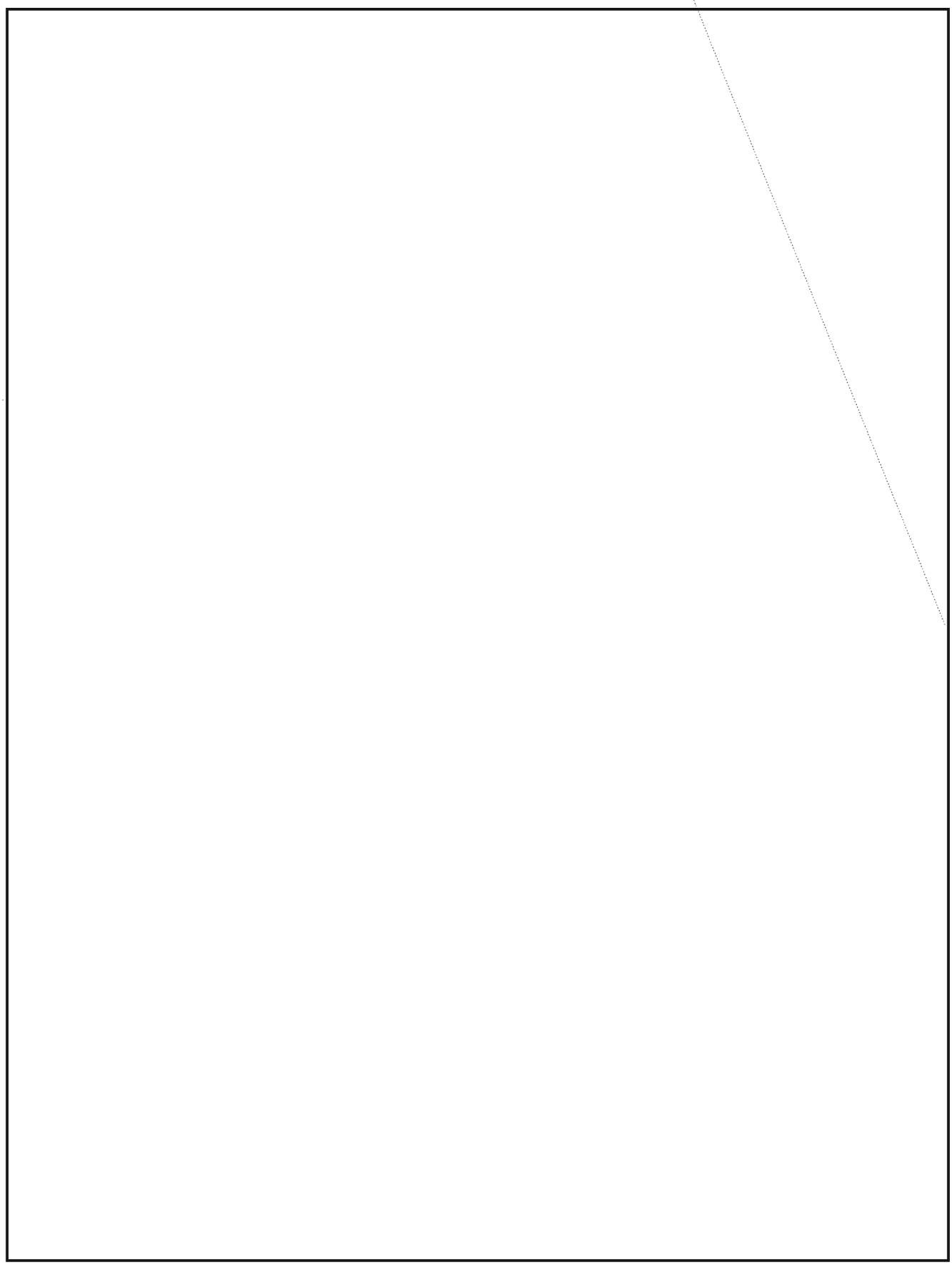
What can it do for us?

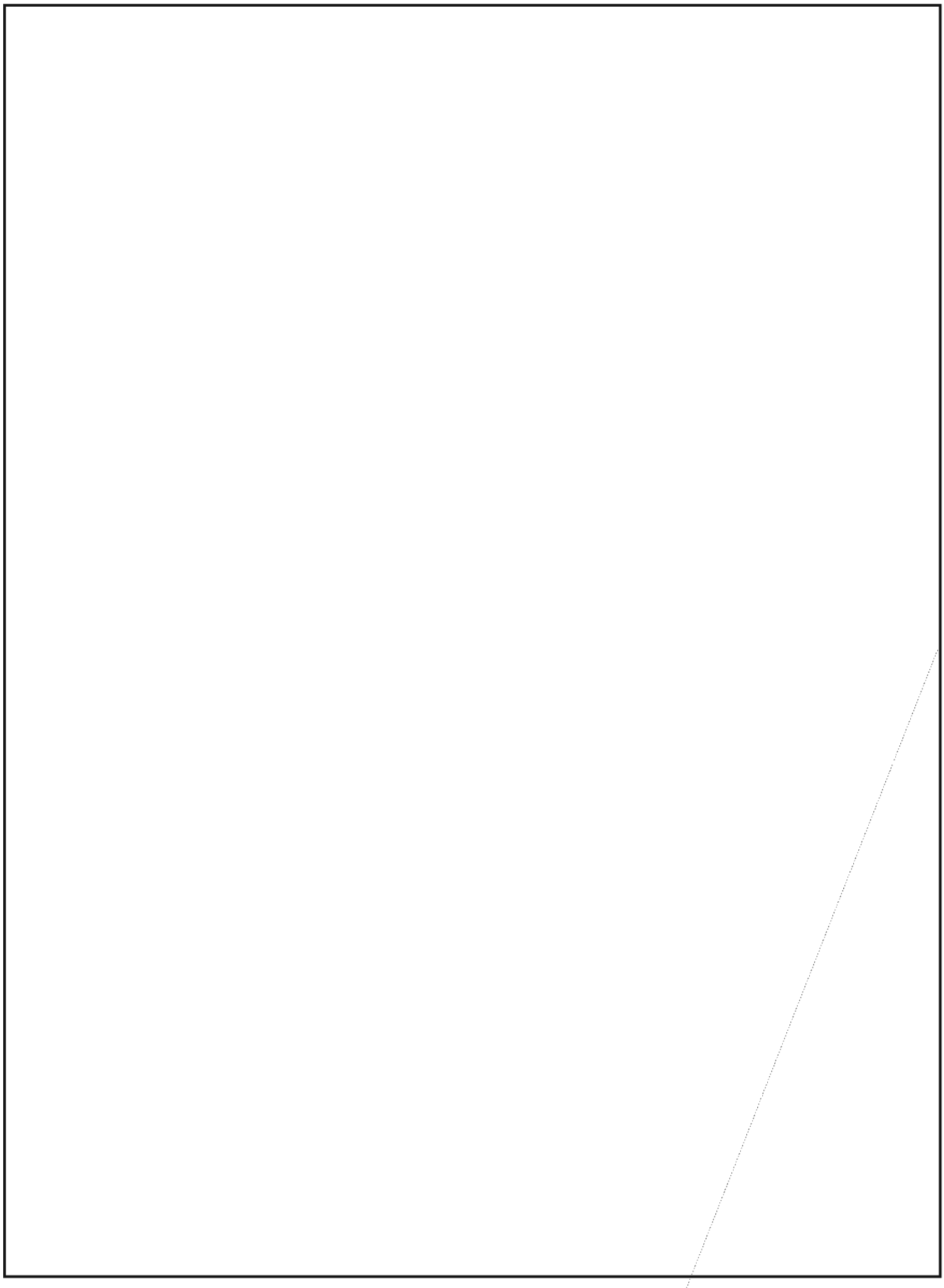
by [redacted] P.L. 86-36

[redacted]

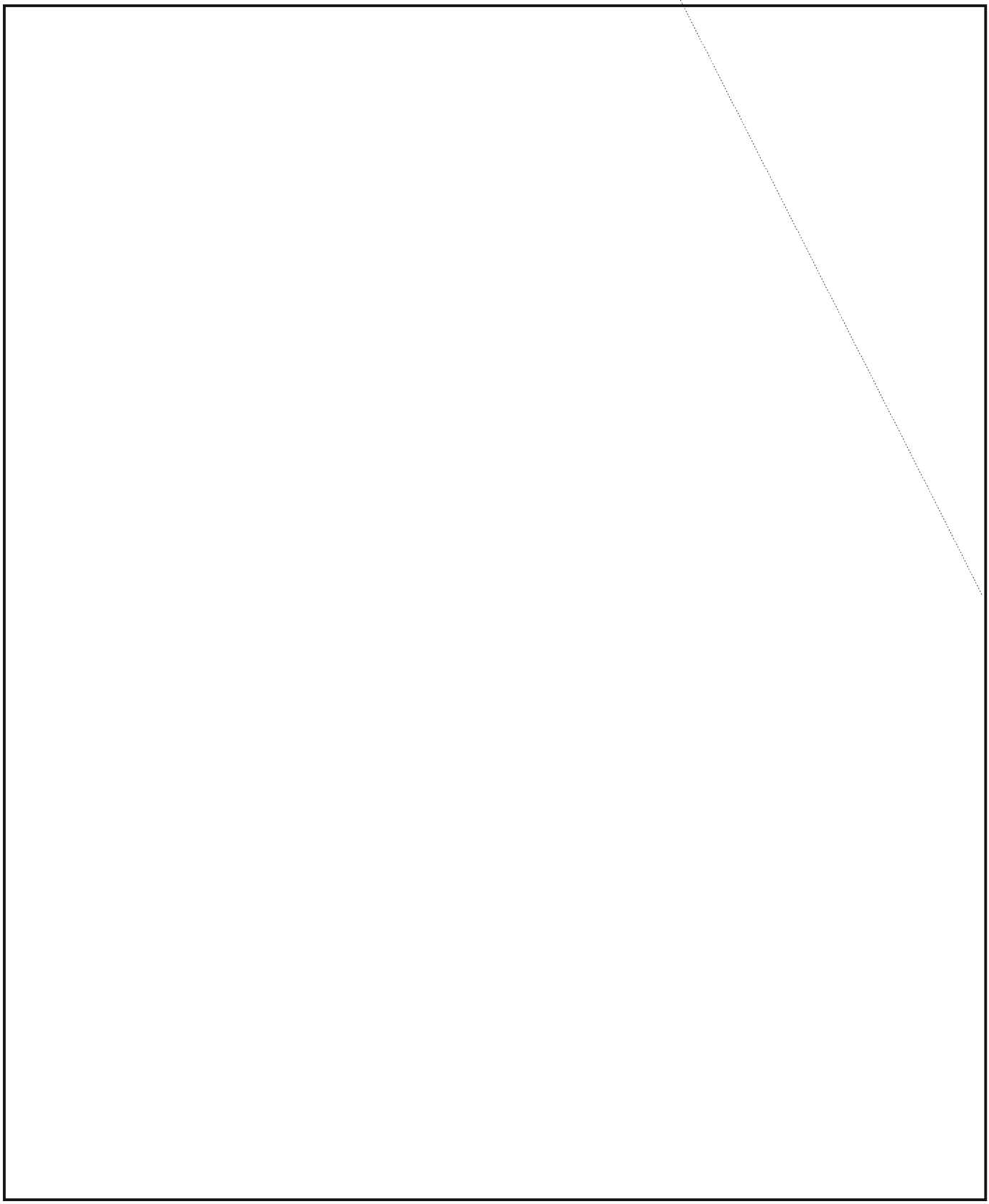
EO 1.4.(c)
P.L. 86-36

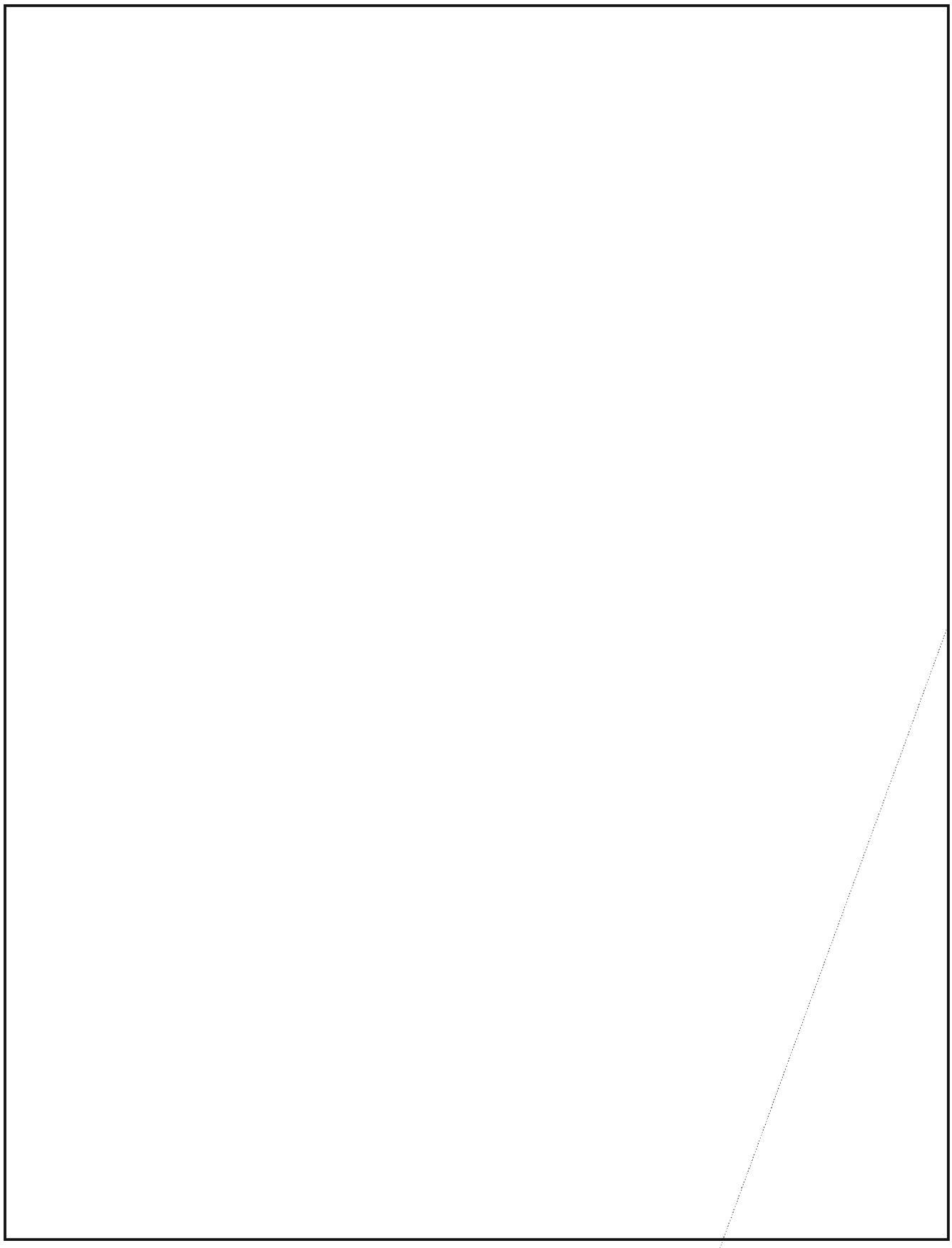
~~TOP SECRET UMBRA~~



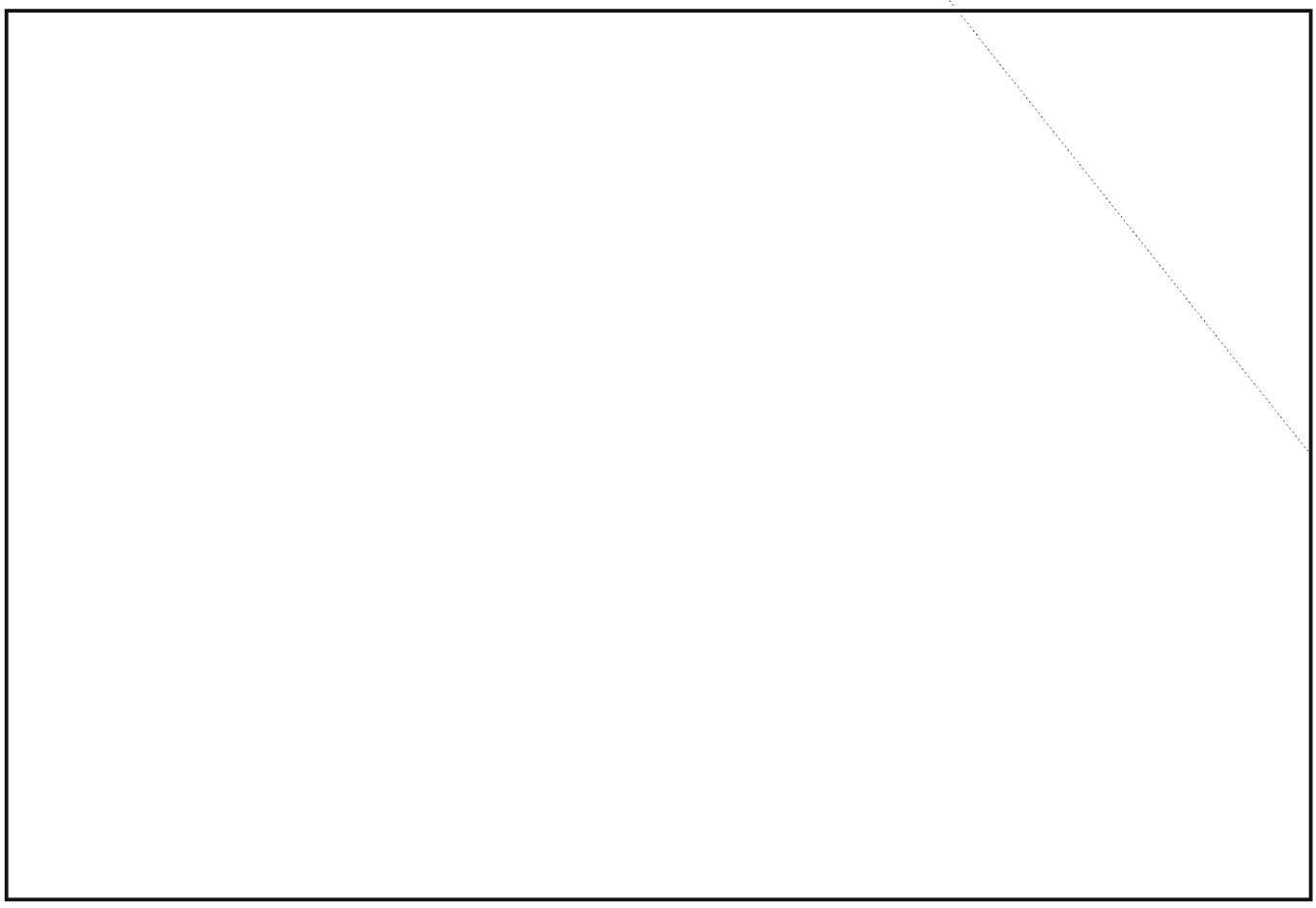


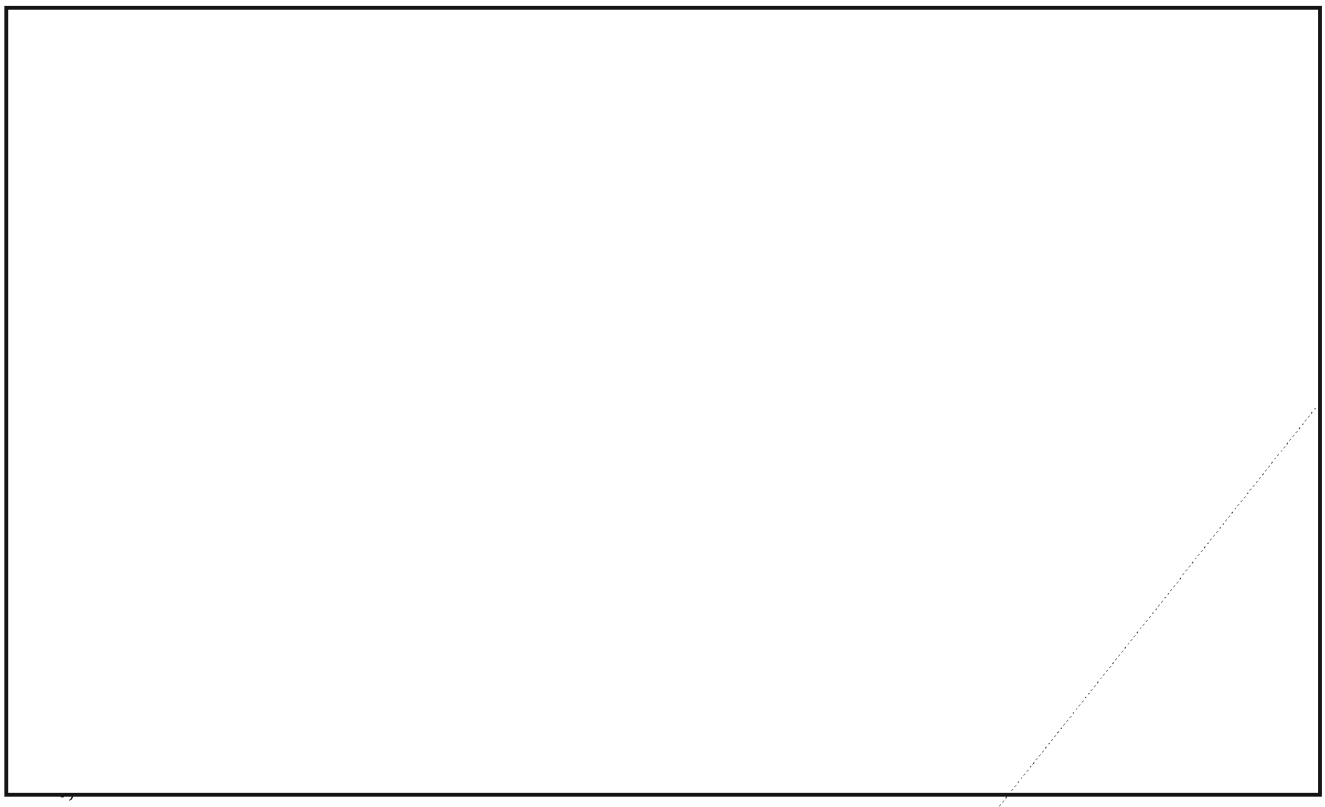
P.L. 86-36
EO 1.4.(c)

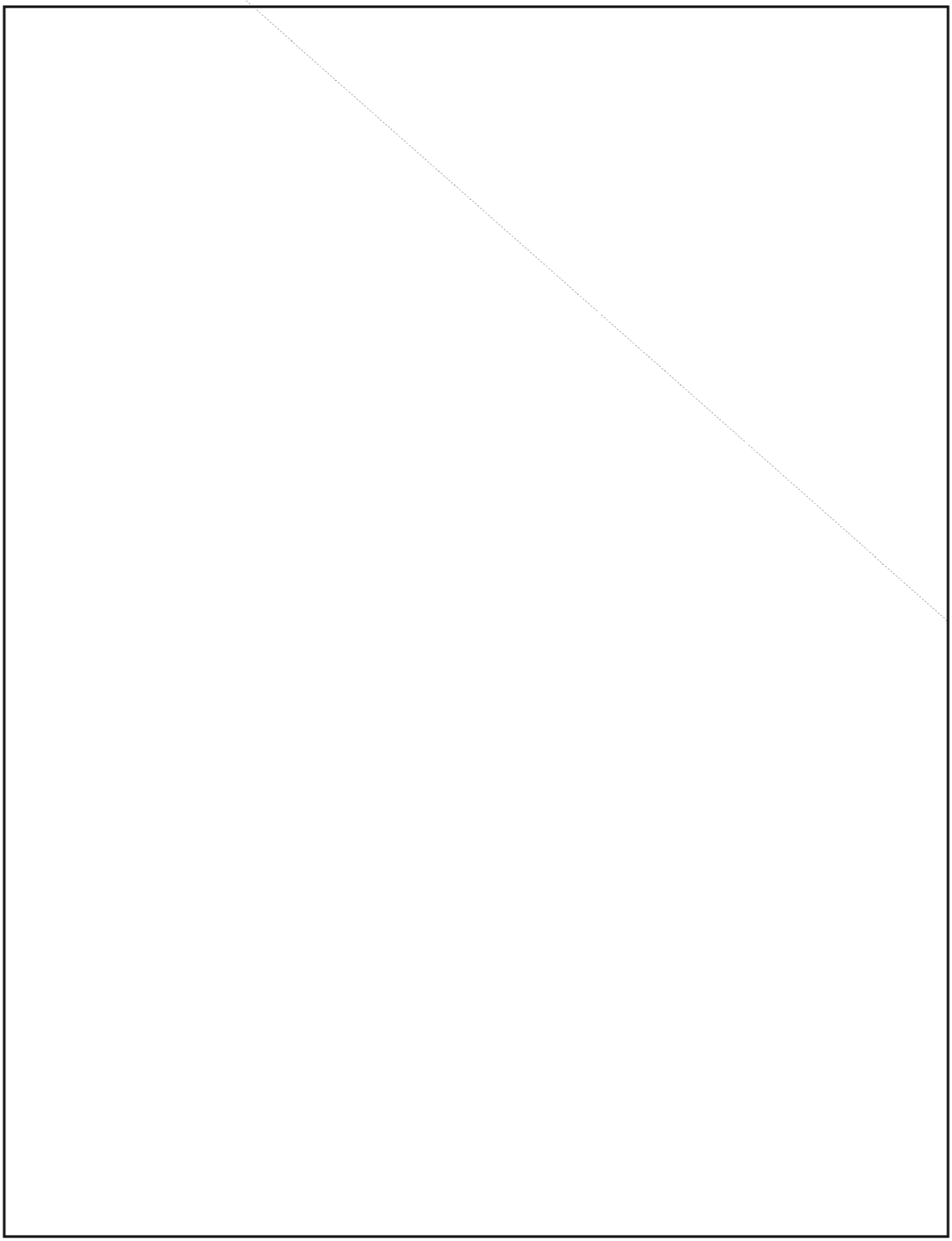


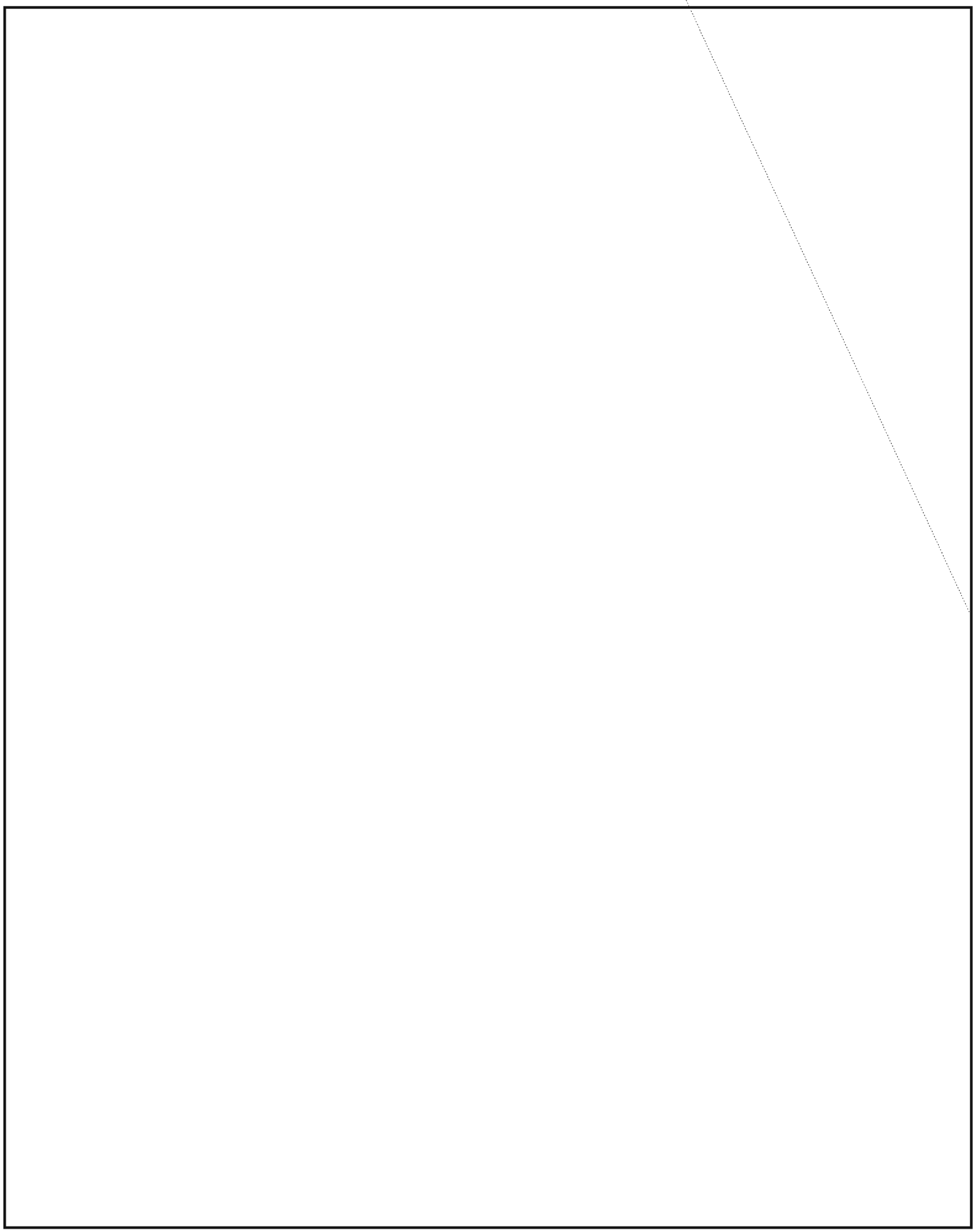


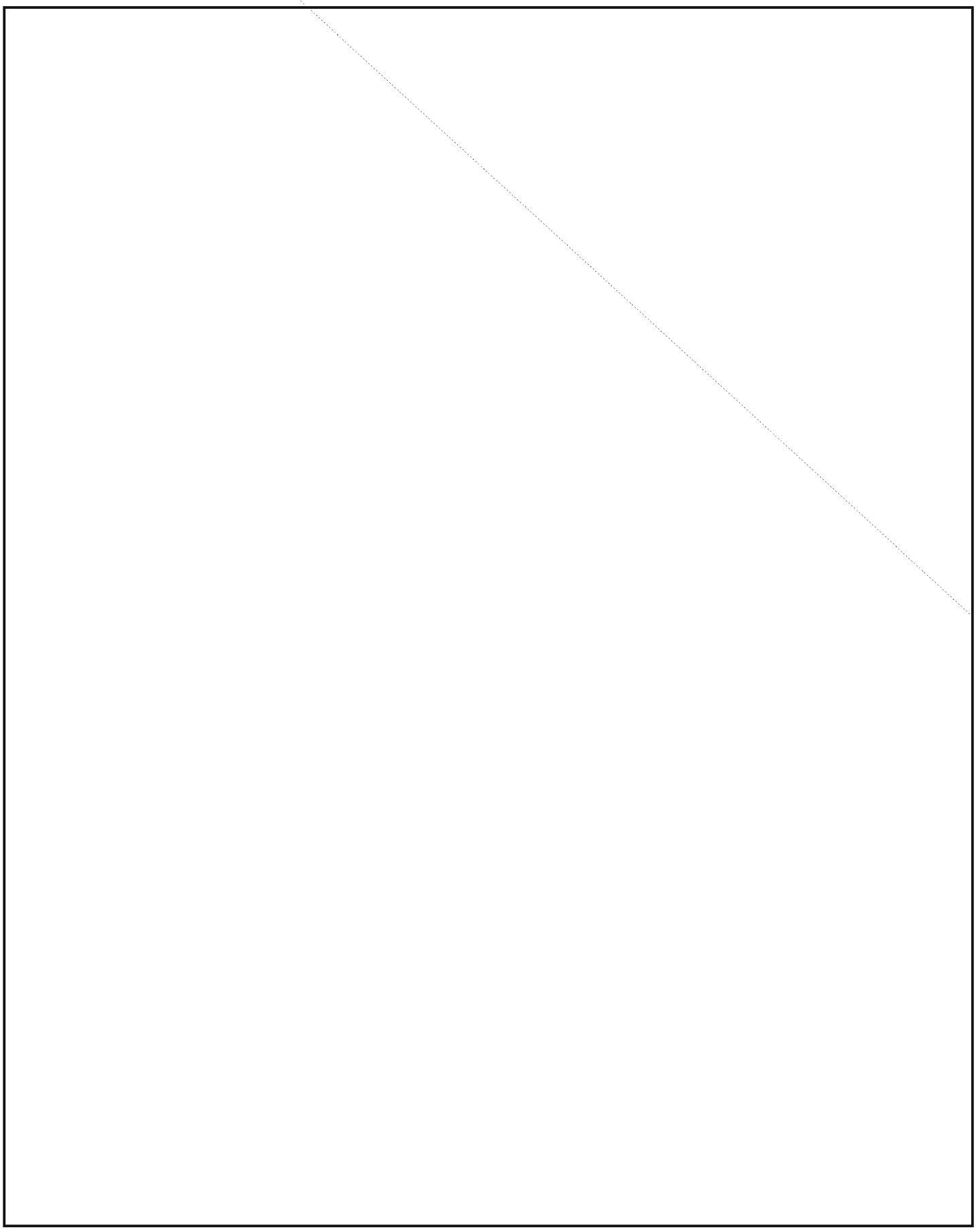
P.L. 86-36
EO 1.4.(c)

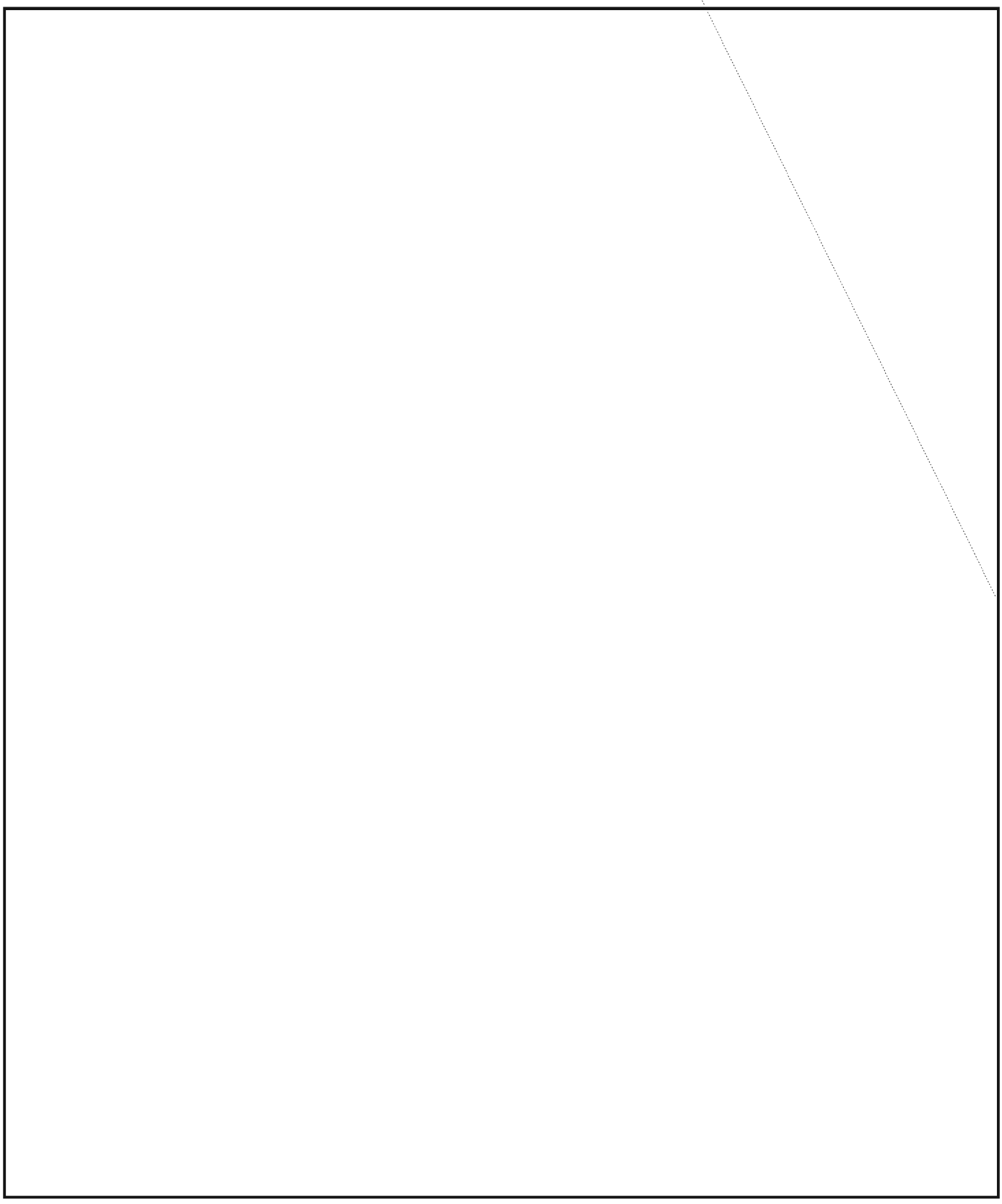


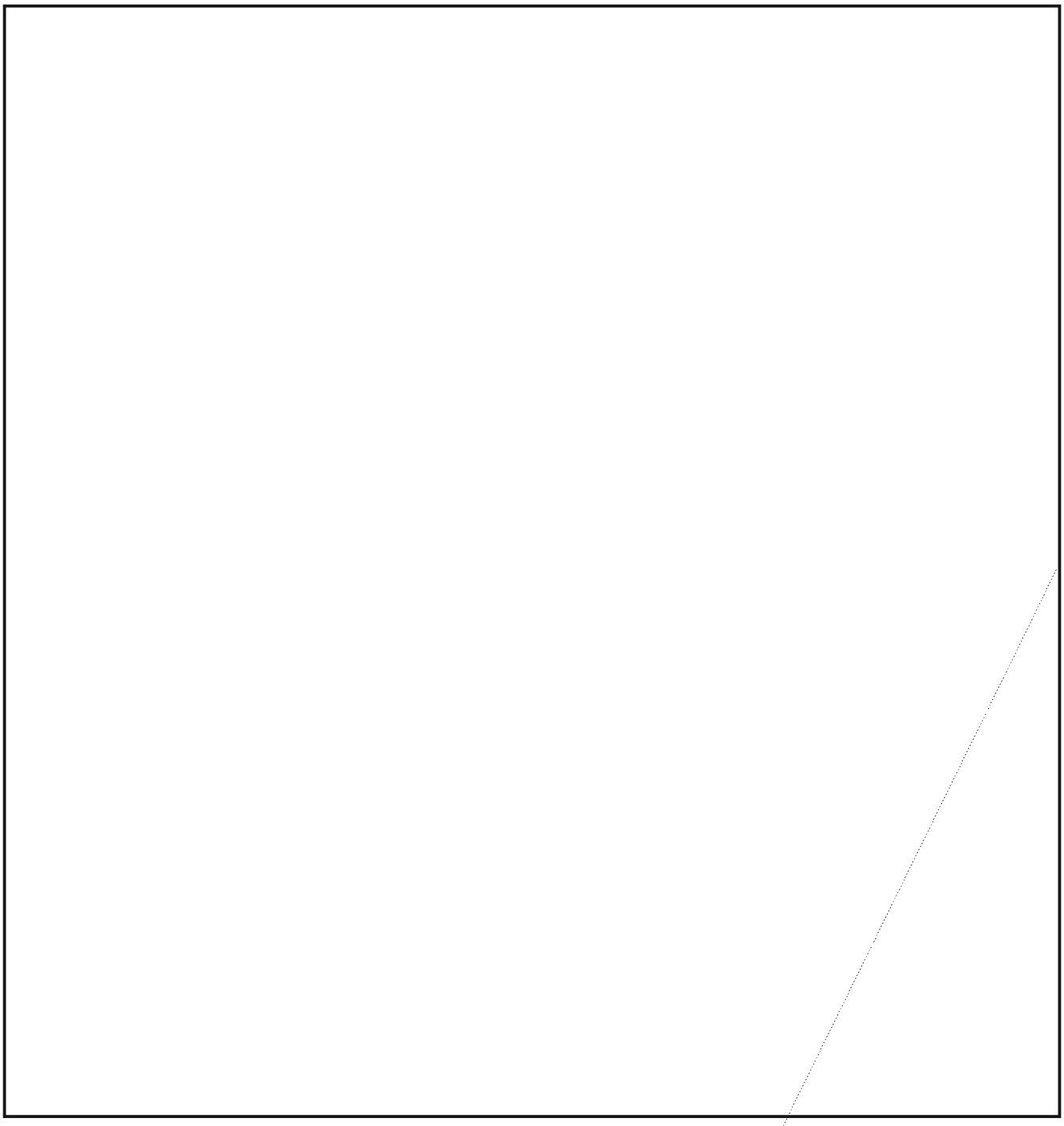












P.L. 86-36
EO 1.4.(c)

CRYPTOLOG
Fall 1995

Software Development:

We Can Do It Better—and Faster

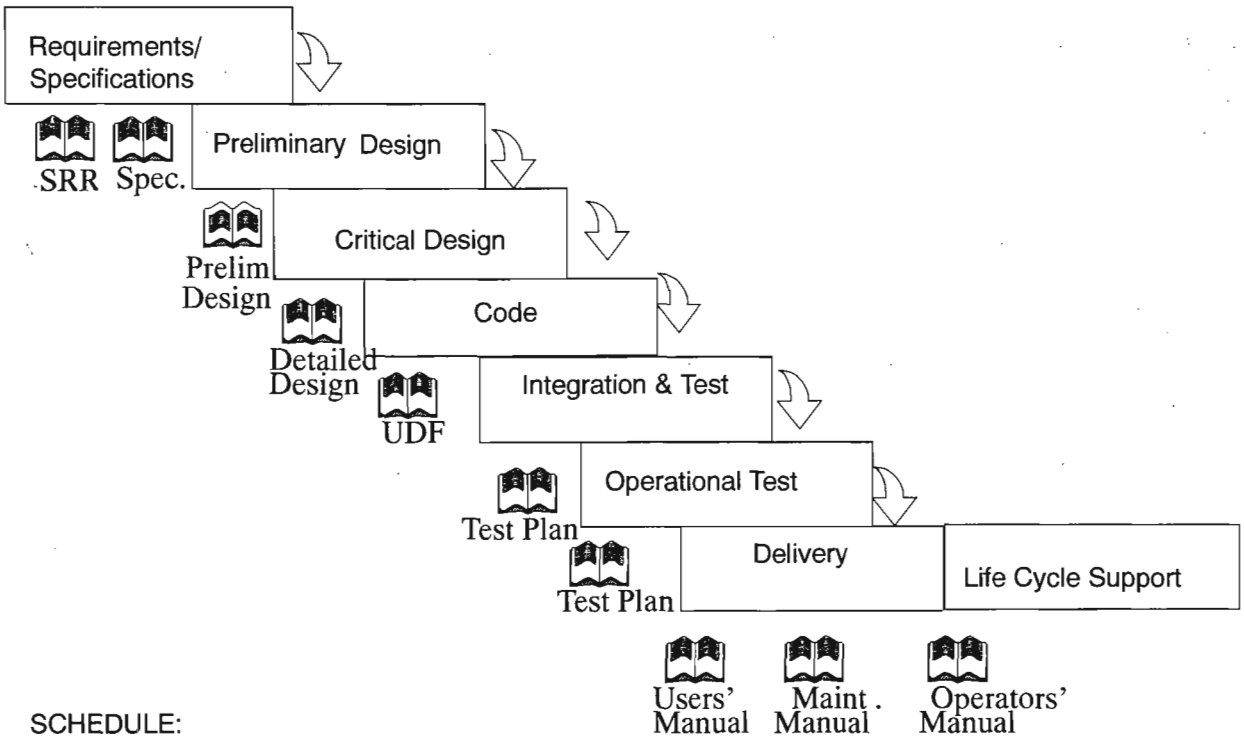
P.L. 86-36

by [redacted]

(U) I believe we can develop better software, deliver it faster and save almost 70% of the cost by changing our acquisition and development methods to what I call "continuous" development. We already do some development this way; we call it maintenance.

(U) In the early days, programming was an art. Programmers were regarded as temperamental artists who produced dense code that was incomprehensible to other programmers, and even themselves after a couple months had elapsed. (As a humorous signature block on the Internet put it, "Real programmers don't document; if it was hard to write, it should be hard to understand.") Development schedules and costs were unpredictable. Code was difficult to maintain. There was no discipline in the process.

(U) Today's software acquisition process, based on a fear of failure, is at the other extreme. Systems analysis and systems engineering principles are applied, and have become disciplines of their own apart from programming. There is great emphasis on reviews and documentation trails to demonstrate that the acquisition manager has done everything that should be done. Each step must be completed, reviewed and approved before proceeding. The goal is to bring order and predictability to the process, to produce code that can be understood and maintained by other programmers, and to produce documentation that will both guide the development and guarantee maintainability over the life of the system. The diagram below shows the essence of this method—often referred to as the Waterfall Method.



(U) A lot of common sense is embedded in these steps. Surely it makes sense to understand and document what you are going to build before you build it. And all the interested parties need to sign up that they agree with the statement of what needs to be built. Likewise, it makes sense to design the system at a top level before proceeding to more detailed levels of design. Like drawing a picture, you can keep the proportions in line if you roughly block out the picture before working on the details of any part of it. If you don't block the design out first, you run the risk of re-creating the old sign we all know and love.

THINK AHEAD

(U) Surely testing is a necessary step. All these steps are necessary. It's plain common sense that disaster would ensue if one of them were omitted. What, then, are the problems?

(U) Let's examine how the process goes astray. The typical acquisition starts out with a year or so of planning activities. Along with starting up the 25-5 paperwork to gain project concurrence and approval, there is a widespread effort to gather all the requirements for the system from the user population. The typical approach to development calls for a set of requirements that are common, consistent, complete and set in concrete. A lot of effort is spent in the process of gathering and coordinating the requirements. Generally, real users do not participate in this process. They are busy doing their jobs, they do not speak "requirement-ese", and they are often cynical about "wasting their time" on a development unlikely to succeed—success being defined as delivering a system the user likes within budget and on schedule. Thus requirements gathering usually is turned over to pseudo-users (user representatives, customers, customer representatives, etc.) who try to specify each capability that will be needed over the life of the system with enough detail that it can be turned into a testable system specification. It can take many months to generate the requirements document. The accepted understanding that a requirement not specified at this time cannot be added later leads to over-specification of the requirements. Requirements that are not fully understood and perhaps not necessary are specified anyway, because of this "now or never" philosophy. When the development runs into trouble, as these developments inevitably do, the requirements process is correctly blamed.

(U) The fix is to realize that the problem was over-specificity, not under-specificity. The development did not fail because some crucial requirement was overlooked. It failed because there were too *many* requirements. There is no good way to sort out the really important requirements from the "nice to haves". In addition, the sheer volume of the number of requirements contributes to the difficulty of understanding them. They are indigestible because of their mass.

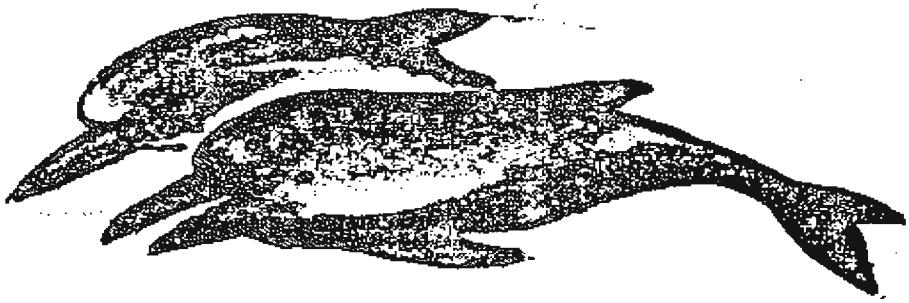
(U) When I was a very junior computer scientist years ago, a co-worker and I were faced with the task of writing the requirements document for a contract. We knew perfectly well what the job was, and could have drawn the top-level design given a moment's notice. Struggle as we might, we could find no way to preserve and communicate our understanding of the problem while using the format required for the Requirements Document. When we finished, the document was incomprehensible even to us, although we had carefully double and triple checked that all the requirements were correct and were included. I think we were both very glad an experienced contractor, who knew how to deal with that mess, was the recipient.

(U) Of course the contractors don't have any magic either, especially when they are unfamiliar with the subject matter of the contract. They have difficulty understanding the big picture of what they need to build and how the system will be used operationally. Imagine trying to assemble a bicycle the night before Christmas without any concept of what a bicycle is—or even the picture of the bicycle on the front of the package. In fact, we religiously keep the picture of the bicycle from the developer because that would imply "a design" and we must give them pure requirements untainted by design assumptions.

(U) Developments usually have a step called System Requirements Review (SRR), which intends to embody the wise practice of "repeat the task back to me so that I can be sure you understood it." Unfortunately, the SRR document that is the medium of communication is as unintelligible as the Requirements document it responds to. Furthermore, the process is hindered by the program review format. Large documents are mailed out to the SRR audience a week or so before the review. Reviewers come to the SRR documents fairly cold and have the monumental, if not impossible, task of comprehending several linear inches of documentation while continuing to perform their other job duties (after all,

they have not just been sitting on their hands doing nothing while the contractor produced this document). The review itself generally consists of several 8-hour days spent presenting viewgraphs of the material in the documents. From the contractor's point of view, the review is a success if no large problems are discovered. Even if they are, there is no time in the schedule for going back to re-do this step, so the plan is to fix the problems later. The result is that after SRR, the developer is free to proceed with the next step of Preliminary Design. And so it goes through the Preliminary Design Review, Detailed Design Review, Test Plan Review. The problem is that the model does not fit how humans think and communicate:

(U) Continuous development consists of breaking the job into small manageable releases. Each release should be a simplified working version of the whole system. We often talk about peeling the onion. Continuous development is like *building* the onion layer by layer. Start with the essential core processes in simplified form and build the framework. Refine and elaborate on that framework in subsequent releases. Each release should accept real data (modify some other data or simulate if you have to) and put out real data in real formats or displays. It should be given to real users to try. If it cannot be used operationally in the early stages, then users should be able to run it for evaluation. My experience is that there are typically four to five releases before the system has all the capabilities originally envisioned. Developer foreknowledge that several more releases are necessary on top of the first one works magic in producing a maintainable design with reduced integration problems. If the design is not maintainable, the developer will learn rapidly on the next release. It is a self-correcting situation. Also, by building the whole system in the first release, they will have had to integrate all the parts. Subsequent deliveries will modify and enhance the already-integrated pieces. You avoid many, many problems by integrating early while the pieces are relatively simple. This is a fundamental strength of continuous development.



"Although humans make sounds with their mouths and occasionally look at each other, there is no solid evidence that they actually communicate among themselves."

(U) The SRR steps *are* necessary, they are just overdone. They are also necessary for the life-cycle-support phase following the development cycle. So, the traditional method actually divides the work into two phases, development and life-cycle support, and maintains the fiction that the life-cycle support phase just fixes minor bugs and keeps up with new releases of the operating system. The constant emotional wrangling at most Configuration Control Boards about whether to call the new work fixes, enhancements, or new requirements should be a clue that reality and the model do not match. There is really one extremely large development cycle followed by numerous smaller development cycles. Why don't we get smart, forget this fiction about "development $\neq$ maintenance", and just develop the whole thing incrementally?

(U) Incremental or continuous development is a model that *does* fit the way humans think and communicate. All of the same steps necessary in the traditional waterfall development are present in continuous development, but in a smaller scale. No one can design without knowing the requirements they are going to address. No one can code without some concept of a design. But the continuous development runs through the waterfall steps for each delivery. There is no separately identifiable maintenance phase, just smaller deliveries as requirements taper off.

(U) The development should try to concentrate on the hard part first. If the hard part is keeping up with a high volume of input data, the developers should concentrate on the parts of the system that deal with the data volume first and go with simplified user graphics until later releases. This is in sharp contrast to most acquisitions, which typically commission studies of the hard parts. For the same amount of time and effort consumed by a paper study or even simulation, you can have real working software that can be run and measured. One can find the bottlenecks and fix them. Even if the software fails (which I have never seen), you learn much more from the software than would have been possible from the study or simulation.

(U) Each release should be built and delivered quickly: in six months or less. This may be the most important rule. It guarantees that the development cannot go too far astray before everyone knows it. No more going directly from "everything is green" to large overruns and delays.

(U) Software is truth; it is not vaporware or shelfware. It either works or it doesn't, and it is there for everyone to see. This is enormously motivating for developers. Anyone can size a six-month effort: a couple of weeks of understanding the requirements and design, maybe four months coding, a month or so of testing and documentation, and a couple of weeks to allow for slips. A pass/fail grade will be delivered before anyone can move on to another job. And people will work incredibly hard and become inspired in their efforts to avoid failure.

(U) *The team must be small, usually four to seven people.* This follows naturally from the small rapid deliveries. You cannot put fifty people on a six-month piece of software. The benefits of the small team are that real communication is possible and each team member has a good understanding of the whole system and how all the pieces fit. This enhances the quality of the design. The system must be designed as a whole and the designers, coders, testers and documenters are the same people. When the increments are small and rapid, one avoids the problems caused by assigning a designer to each function and producing a system that looks like it was designed by a committee.

(U) Another factor is that, given the rapid nature of the development, there is no need to try to communicate through large design reviews and multi-inch documents. Instead of having formal reviews of documents that several layers of management have pre-reviewed, just review the software. Run it. Measure it. Software is truth, while documents can obscure truth. And it saves a lot of useless work and money/manpower for both the customer and developer side. Today it is common to have key developers spend a whole month preparing for a review instead of working on your system!

(U) The final aspect of continuous development is user satisfaction. Real users—the "stuckees"—are involved in the process because you speak to them in their language. You are showing them the real system as it progresses, allowing them the opportunity to influence future releases, and doing it all in a reasonable time frame.

(U) I like to compare building software to building a house. The traditional waterfall method would have us lay out all the requirements needed to make the house a turn-key operation: furniture, lights and rugs in place, towels folded to spec in the linen closet, curtains on all the windows. The house would be divided into subsystems, with lead designers and programming teams for each. Each team of five to ten people would then work on their detailed subsystem design, code, and perform unit test. Most real problems would not be evident until integration is attempted and the linen closet will not fit into the bathroom. The way houses are built in real life, and the way software should be built, is to put down the foundation and some framing first. Framing equates to the first incremental delivery. The user can walk through the house and decide that the traffic pattern through the kitchen is wrong and a door should be moved. Windows, and even staircases, can be relocated. It may take a little extra time and effort to move things at this point, but a better system will result. The alternative is that the user doesn't get to see how the traffic patterns, doors, and windows fit until integration or delivery, and then it is too late to change. They are stuck.

(U) I have discussed this over the years with various skeptics, and a frequent response is that this concept is all right for small analytic systems, but it does not apply to large automated systems. It is true that this type of interactive, iterative development is ideally suited to in-house analytic systems; that *is* how some of our most popular analytic tools (OILSTOCK, TINMAN, SCREENWORK, SUNSHINE, etc.) have been developed, some by small contractor teams and some by small government teams.

(U) However, I have personally used this method of development twice to replace large multi-year, multi-million dollar semi-automated processing systems. The first was the FDPS at Sunnyvale, and the second was

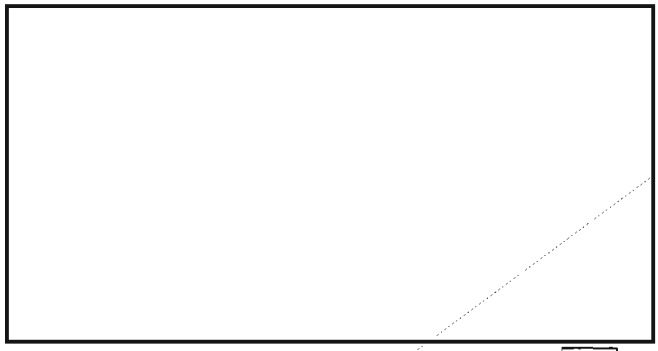
MINSTREL. In both cases, we started an alternative development long after the acquisition was underway because we were convinced that the acquisition system would not operate satisfactorily. In the case of MINSTREL, cost and schedule overruns were also an issue. In both cases, we delivered a much better system earlier, for less than 10% of the cost. I am confident that this method will scale to the largest acquisitions that NSA could conceivably undertake.

(U) Another interesting question I get is, "how do you know when you are done?" Maybe you are never done until it is time to replace the system. As long as there are users, there will be new requirements. Satisfying those new requirements will make the users more productive, and that is why we have ADP support. Since the continuous development team is smaller than many maintenance teams to begin with and since there is a maintenance team for the life of the system even for waterfall acquisitions, I am sure there is not a problem recognizing that continuous development releases and maintenance releases are actually the same thing.

(U) Documentation is another point. Just because you have reduced unneeded documentation doesn't mean there is no documentation at all. The requirements for each release are negotiated between developers and users, and documented to provide guidance and reduce misunderstandings. The code is documented and commented as necessary; since insufficient documentation in one release will cause serious problems in subsequent releases, the developer has more than usual interest in providing adequate design and code documentation. The users' document will improve with each release, especially as user input is incorporated. Finally, documentation for system administrators will also improve as each release is installed.

(U) Continuous development should be adopted as the standard NSA way of doing acquisition. Mil-Std 498, which replaces the NSA 81-3 standard, has the framework for continuous development built in and encourages its adoption.

(U) In summary, continuous development delivers useful results within monthst is much cheaper because fewer people are needed: six or so, versus sixty to one hundred. It eliminates unneeded documentation and the expense of massive program reviews (both man-hours and presentation graphics). The development is a satisfying experience for both the development team and the users. Above all, it works.



Kλ

P.L. 86-36

Foreign Language Testing at NSA: Time for Change

by [redacted] H431

(U) This article reports the findings of a study carried out in mid-1995 touching on certain aspects of NSA's language testing practices. As an NSA fellowship recipient, I undertook this study in partial fulfillment of requirements for a Ph.D. degree in applied linguistics from Georgetown University, and in the hope of provoking beneficial changes to the current language testing system. It was conducted with the permission of M09 and the Language Career Panel.

(U) I expect that readers of this report will include both experts and nonexperts. For more on the theoretical background, full results, or procedural details, my dissertation (now in progress) will be available. (Many of the comments in this report are intended for NSA only and will not appear in the dissertation.) This article identifies two major threats to the fairness of NSA's language testing which should be addressed immediately; various other potential problems of a less serious nature are also pointed out and recommendations are made as to how they might be corrected. These comments are meant as a starting point for discussion and not as the definitive answer to all of our testing problems.

Objectives

~~(FOUO)~~ I had two main goals in this research, as far as this agency is concerned. First and foremost, I was interested in establishing the validity of the method that NSA and other government agencies use for choosing foreign language test passages. *In a nutshell, I wanted to know whether level 2 texts really are easier to comprehend than level 3 texts, and level 3s easier than 4s, for test takers at all levels of proficiency.* This must be true in order to claim that the levels scheme can appropriately be used as the basis for our testing system.

(U) Furthermore, even if one assumes the validity of the text levels, it is clear that there is a need to increase the reliability, validity, and efficiency of our foreign language tests. Hardly a language analyst has not complained about some aspect of the PQEs; most of them have a feeling that something is wrong, even if they cannot say exactly what. On the basis of my experience in this study, I make several suggestions here for consideration by the appropriate NSA elements.

(U) The study involved testing 56 employees with French language backgrounds for their French reading proficiency. I chose French because, frankly, that is the one foreign language that I know well enough for designing a good-quality test with a high degree of difficulty; and I had to leave aside the area of listening comprehension because of time constraints. Although the results reported here thus can only truly be said to apply to French reading comprehension, it would surely be necessary to show that the testing system works for even one language before we could properly try to extend it to all. Likewise, if the system is not applicable to reading (and by extension, to translation), then applying it to listening comprehension would not likely be fruitful.

(U) Rather than describing the various text levels here, I assume some familiarity with them on the part of readers. I must point out, however, that it is a common mistake to oversimplify what a given text level means, and thus I encourage readers to consult James R. Child's 1987 paper (which is included in the materials for the self-paced course LG-020, "Language Levels and Their Application") for a full description. I am completely ignoring the Interagency Language Roundtable's (ILR) levels 1 and 5, which is where the ILR setup runs into some real theoretical problems; these levels are not of any practical concern to government agencies anyway. In addition, be advised that on theoretical grounds, I do not accept the reading skills hierarchy that the ILR scale incorporates.

Part I: Text levels

Reliability

(U) *The business of determining the level of a text involves judgements by human beings.* Now, we must be fairly certain that, for instance, a level 3 passage really is a 3 and not a 2+ or a 3+, if we are designing a

level 3 test, because we want to ensure a consistent level of test difficulty. In fact, if text levels cannot be reliably determined even by trained experts, there is no sense in going any further in evaluating the validity of the levels theory.

(U) Just as an advertisement that says "nine out of ten dentists recommend Brand X" tends to inspire more confidence in Brand X, we tend to have more confidence in text-level judgements that are backed up by many expert raters. In the ideal situation, if we gave three or more experts the same set of written passages and asked them to determine their levels, the experts would—working completely independently—come up with exactly the same level assignment for each text. In the real world, we have to admit that people are not infallible and that it is perhaps harder to get linguists to agree on text levels than it is to get dentists to agree on chewing gum, but we should still strive to get as close as possible to that ideal target of complete agreement.

~~(FOUO)~~ In developing my experimental materials, I first identified a large number of authentic (naturally occurring, uncontrived) French texts between 250 and 300 words long, at text levels 2 through 4, including some that I thought were 2+ or 3+. I took 30 of these to others for independent decisions about levels; each text was rated by three people (including me). My experts were all current or former PQE committee members who had completed LG-020.

(U) The good news is that almost 97% of the time (29/30 cases), it was possible to get at least a two-way match (at least two experts agreed). Unfortunately, only 23% (7/30) were three-way matches, which is the ideal.

There are two things we can do, given this less-than-perfect situation. One is as follows:

- **Suggestion 1:** Use, in tests, only those texts for which three or more experts independently agree on the level.

(U) In selecting passages to use in this experiment, I gave preference to the three-way matches, using five of them in my test (which contained nine passages in all).The other thing we can do is to try to increase the number of three-way matches by improving the expertise of the raters. One way to accomplish this would be to make it a little more difficult to qualify as an expert rater. This could be done as follows:

- **Suggestion 2:** Double (or even triple) the number of items on the LG-020 exit exam (thereby increasing its reliability), then adopt a higher standard of performance, of 85 or 90 percent.

Another way to promote greater agreement is this:

- **Suggestion 3:** Require test designers (such as PQE committee members) to "socialize" at the start of each testing cycle, discussing several texts that are in the pertinent foreign language and that have previously been determined to be at the various levels.

(U) "Socializing" in this sense is obviously not what happens at cocktail parties, but it means making sure that everyone is interpreting the guidelines in the same way, often by studying examples.

(U) In making the suggestions in this section, I do not wish to imply that we necessarily have a serious problem right now with text level reliability. Reliability is difficult to estimate in this situation, and of course we can always try to improve it. Nonetheless, we may in fact have all the right ingredients for an acceptable level of reliability, *provided that the proper procedures are followed.*

Validity

(U) Now, just because we can determine text levels reliably does not mean that the levels do what we think they do for us. Reliability is a *necessary* condition for validity, but it is not *sufficient*. Let us consider an example of how something can be done reliably but still not be valid. Suppose I decide that all magazine articles are easier to understand than all newspaper articles, and the latter are easier to read than all books. I have no doubt that almost anyone could tell a magazine from a newspaper from a book with a very high degree of reliability, but few people would really be convinced that I have identified a true progression of text difficulty.

(U) It is obvious in this example that something is wrong, but it is not always so easy to determine this. Someone who has read Child's paper might say that the text-level scheme *seems* right for estimating text difficulty. Intuitions are useful for many endeavors, but sometimes they, too, are dead wrong. Science is full of examples of nature contradicting our intuitions (isn't the earth flat?). That is why we often want experimental data to confirm (or refute) what intuition tells us.

(U) It has never been shown experimentally that texts at the different levels really define a hierarchy of

comprehensibility that can be used to separate test takers by level of proficiency. In fact, researchers in academia claim to have found evidence in two widely cited studies that this is not the case. These earlier studies, however, were seriously flawed, suffering from the following problems:

a. The researchers reduced the text-level descriptions to simple genre labels, so that any editorial, for instance, was taken as indicative of level 3 with no further analysis. Level decisions seem to have been based solely on genre, with no concern for the communicative functions that the texts served.

b. Only one or two texts of each type were used, and these texts were not chosen by multiple trained, independent raters. The reliability of level assignments is thus suspect.

c. The range of foreign language proficiency of their subjects did not begin to cover the full ILR range. The higher levels of ability were particularly underrepresented.

~~(FOUO)~~ I set out to conduct a similar experiment that would remedy these problems. I designed a test composed of nine texts of about equal length, including three at each of the levels 2, 3, and 4 (avoiding any "plus level" texts for maximum separation of level effects). The texts were on various topics, including political affairs, social affairs, terrorism, and human rights. Potential test-takers were randomly chosen from among all those who had passed a French test at the Agency within the past 15 years. This included many people who had never used French on the job as well as certified French language analysts, so a wide range of ability was represented.

(U) In the testing sessions, each of the subjects saw multiple-choice questions on six of the reading passages, and did "rough translations" (described in more detail below) on the other three passages. Test versions were rotated so that each text was translated by about one-third of the subjects; the order of presentation of the passages was also varied to control for warm-up and tiring effects.

(U) A 27-item multiple-choice French reading comprehension exam taken from tests designed by the University of Ottawa was also administered to all subjects. This independent measure of proficiency was originally supposed to allow division of the subjects into three groups by ability, but it was not reliable enough for this purpose. Thus a combination of scores on the two multiple-choice sections (Ottawa and experimental) was used to divide test-takers into three groups of higher, average, and lower ability, and my analysis then focused on the translation data only (scoring of the translations will be discussed in Part II below).

(U) As mentioned earlier, for the theory to be valid, the level 2 texts had to be easier for everyone to translate than the level 3 texts, and the 3s easier than the 4s, in a statistically significant way. My analysis of the translation data shows that this is indeed the case *on average*; there were significant differences between the means for levels 2, 3, and 4.

Table 1:

Text set	Level 2	Level 3	Level 4
terrorism/ human rights	87.2	84.7	70.0
social affairs	81.3	83.4	81.5
political affairs	93.3	76.6	68.5
Total	87.3	81.9	73.3

Mean scores on translations (expressed as percentages)

(U) In addition, the most able group of test takers (independently determined by multiple-choice test scores) had to get significantly higher scores than the least able group, particularly on the level 3 and 4 texts. This was also the case, as the following data demonstrate. (The average group was not sufficiently differentiated from the high ability group. This may be due to various factors, including the mediocre reliability of the criterion used to divide subjects by ability, or the different aspects of linguistic competence tapped by multiple choice and translation tests.)

Table 2:

	Level 2	Level 3	Level 4	Total
High	90.5	87.3	80.6	86.1
Middle	88.8	82.5	75.5	82.2
Low	82.9	76.1	63.6	74.2

Test performance by text level and subject ability level

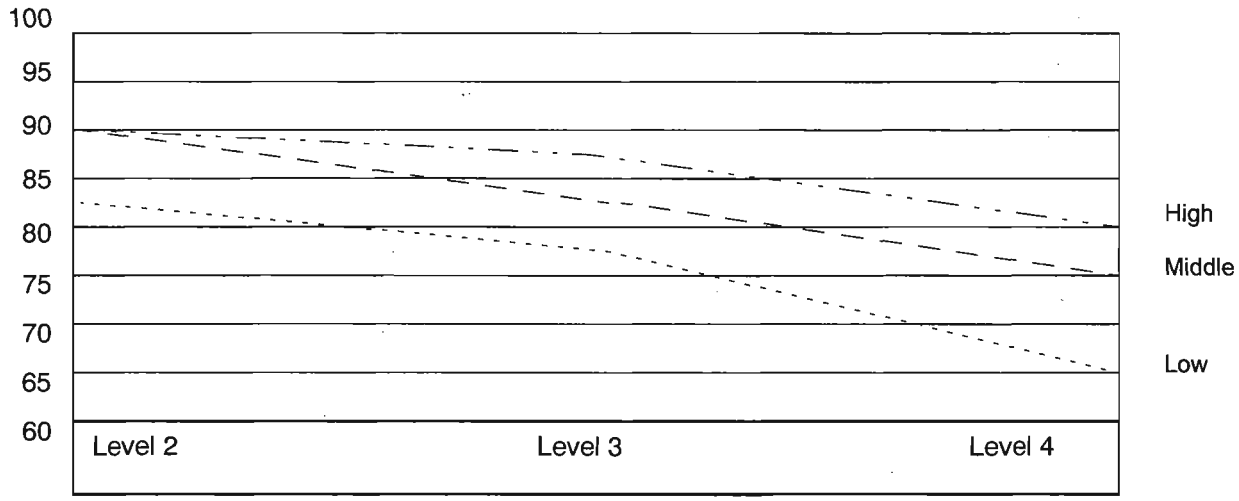


Figure 1. Test performance by text level and subject-ability level.

(U) *One caveat is very important: it is only when a given level of difficulty is represented by* ***at least three texts*** *that the clear pattern emerges.* Group performance on any one text may vary from expectations. As you can see in Table 1, one level 2 passage and one level 4 passage both had about the same mean score, just above 81%—individually they behaved more like the average level 3 text! Each subject in this experiment translated one text at each level; this is supposed to make all test versions equivalent, but in fact, one of the three versions was significantly easier than the others because one or more of the texts in it was easier than predicted.

(U) The need for three texts could be due to some inherent unreliability in the text selection process (see preceding section). It is more likely, however, that the text levels simply cannot be conceived of as distinct entities having relatively well delineated boundaries between them, as in figure 2 on the left. Instead, it may be more appropriate to view the levels as highly overlapping ranges with distinctly different midpoints, as in figure 2 on the right. The levels seem to identify difficulty *tendencies* rather than absolute values.

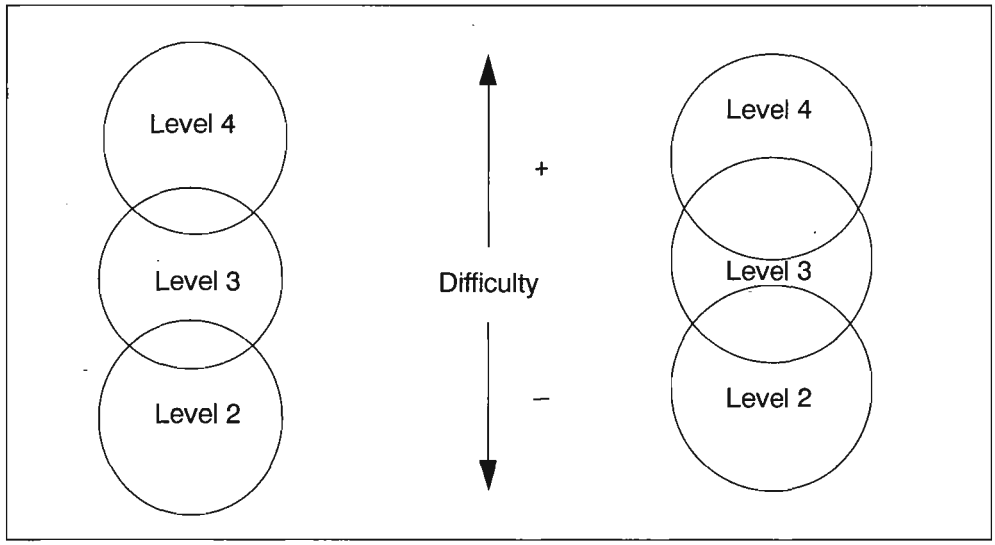


Figure 2. Text levels plotted on a scale of comprehensibility.

~~(FOUO)~~ In practical terms, this means that in order to ensure that test difficulty remains more or less constant from one test to another (as close as possible to the midpoint of the range) without the luxury of pretesting passages to determine their difficulty, we must look at average performance across several texts at the targeted level. This points up the **first major threat** to fairness in our testing: *many PQEs are composed of only one or two passages, so we cannot guarantee that they are all of about equal difficulty.* It is conceivable that we have allowed some people to pass who should not have, and failed some who should have passed. The results of this study indicate that those who are not quite capable of level 3 performance but do not wish to do anything to improve their skills are justified in thinking that one day an easier PQE may come along. Likewise, the capable but not exceptional performer may be unfairly penalized by a PQE that is too hard, being forced to wait until one of more appropriate difficulty is presented.

This naturally leads to the following:

- **Suggestion 4:** All language comprehension tests should employ at least three same-level texts of approximately equal length, from different sources and on different topics.

(U) This recommendation can be implemented immediately without placing an undue burden on test designers. I therefore urge most strongly that it be adopted without delay.

Some have been waiting for an easier PQE.

(U) Equal length is stipulated so that each of the three (or more) texts contributes about equally to the final score. Sources and topics must be varied to ensure that these also are not factors (or at least not important factors) in the outcome. As for the optimal number of texts to use, while clearly more is better in achieving representativeness, there are practical limits to what test takers and scorers can handle, and beyond a certain number, very little would be gained with each additional text. Three is the minimum, but more than five or six would probably be overdoing it.

(U) There is another reason to base tests on multiple passages: by using only one, we cannot be certain that we have obtained a representative sample of each test taker's ability to perform at that level. It is widely accepted in the field of language testing that one should give test-takers as many "fresh starts" as is feasible. Coupled with the findings reported here, this is all the more reason to incorporate Suggestion 4 into NSA testing practice.

(U) A note about "plus levels" (2+, 3+): it is not clear how test takers would perform on "plus level" texts. In terms of text difficulty, does a 2+ behave more like a 2 or more like a 3, or right in between? Most of us probably imagine it as the halfway point between levels, but this is an untested assumption and must be considered suspect. In the absence of any hard data, I would not advocate gearing comprehension tests to "plus levels."

Part II: Other testing considerations

(U) The preceding section may have frightened those who took it to mean that PQEs would now have to be three times as long as before. Rest assured that this is not the case; there are other ways to ensure that language tests are accurate and appropriate!

Test format

~~(FOUO)~~ One might well wonder why I did not use a PQE-style test in my experiment. The choice of testing format, especially for PQEs, has (rightly) been driven by the need to evaluate objectively how well language analysts are equipped to do perform their jobs. Increasingly, however, as the language analysis field changes, analysts can be heard to say that what happens in the PQEs "is not what I do." The traditional translation test is more and more seen as an invalid measure of job competence.

~~(FOUO)~~ *The reality today is that language analysts do a variety of things.* Some must read large volumes of material quickly, making decisions about what they read; some must translate while others simply gist or move straight to an English language report; some

have a great deal of time to complete their tasks, while others work within strict time constraints. Many analysts will find themselves in all of these situations at different times. In designing tests that will indicate competence for this variety of language tasks, we need to ask ourselves what they all have in common. The obvious answer is that a strong ability to comprehend the foreign language is necessary if the language analyst—any language analyst—is to do his/her job efficiently and effectively. That is, language proficiency is the foundation for success in any language-related position in government. Language proficiency, then, ought to be what our language testing is about.

~~(FOUO)~~ The current PQEs place a premium on "idiomatic English" and allow test takers a great deal of time to achieve it. The ability to express oneself well in English is probably much more important a factor in these tests than in most comprehension tests. This is at odds with the way many test users (e.g., supervisors) sometimes see the tests, namely as diagnostic tools for linguistic competence. The hapless analyst who fails her German PQE is thus signed up for yet another course designed to help her understand German, when what she needs may be a course in how to express herself more effectively in English. Unfortunately, the current format has little diagnostic value, as we can never be certain whether poor test performance was due to deficiencies in foreign language comprehension, in English expression, or both.

(U) Now, good English writing ability is no doubt a desired quality in language analysts, and we should probably encourage its development by testing it. Keep in mind that we are not limited to a single testing format. It would be possible, for instance, to proceed as follows:

- **Suggestion 5:** Emphasize speed and comprehension in one part of the PQE, while stressing precision of expression in the other.

~~(FOUO)~~ Since we already have a well-developed translation format in place, I was interested in developing a test oriented more toward measuring reading comprehension. The format I chose for this study, which is described below, is one that allows evaluation of foreign language reading proficiency without unduly relying on English writing skills (as do our current PQEs) or on raw reasoning ability (as do many multiple-choice tests). It is similar to the widely used immediate-recall protocol, but it does not tax the memory and is a more realistic communicative task.

Rough Translation

~~(FOUO)~~ In what I call the "rough translation" format, test-takers are given a limited amount of time to produce a written representation of the meaning of a set of foreign language texts. In this experiment, test-takers had only about 20 minutes to complete a rough translation on each 250-300 word text (or one hour for just over 800 words). This strict time constraint was imposed because researchers have often noted that reading speed is related to reading success. Contrast this with the PQEs, which allow testees several hours to decipher fewer than 600 words, and it is not unreasonable to believe that some nimble dictionary users have been able to pass tests in languages they would not normally be said to "know." (A level 3 reading exam requirement would help to stamp out recreational collection of language certificates.)

(U) Test-takers in this experiment were not allowed to use dictionaries. We may wish to permit dictionary use in actual testing just to reduce test-takers' overall anxiety; however, they should be forewarned that research shows that such use may not help and may even hurt their test scores in a reading test situation.

~~(FOUO)~~ The first step in scoring involves dividing each original text into a countable number of scorable units. This is to provide a meaningful basis for comparison of test takers both within a single test administration as well as across administrations and across languages. This addresses a serious shortcoming of our current scoring system, which is the **second major threat** to the integrity of our translation tests: we have no easy-to-understand, reliable way to determine how many points a text is worth, so we cannot convert scores to a figure (such as a percentage) that can be compared across test administrations. (A method has been recommended based on the count of "propositions" per text, with a maximum score of 8 points per "proposition"; but for some reason—a lack of clarity, a lack of credibility, or a general failure to realize its importance—this method has not been followed consistently.) The result is a points-deducted score that cannot be meaningfully compared to any other scores, except within that particular test. We often say that a test-taker has had *n* points deducted—without saying how many points were *possible*. Choosing an arbitrary maximum number of points for each text—such as 100, as has recently been advocated—does not make the resultant score any more meaningful. A points-deducted method presupposes that all PQEs are exactly equal in difficulty, an assumption that is extremely doubtful (see Part I). One must be able to specify how many points it is possible to obtain

before it will be useful to know how many have been lost.

(U) My model for breaking down the texts into scoring units was the "pausal unit" method, in which several expert readers of the foreign language determine each spot in the original text where it would be possible to pause for emphasis or to take a breath. Everything between possible pauses constitutes a unit. The "pausal unit" approach could easily be implemented by PQE committees with little training. Using my method, most scoring units were between one and four words long. The nine texts used in the study were in this way divided into between 98 and 153 units.

(U) Each scoring unit in which the test-taker had substantially preserved the meaning of the original, without omissions or extraneous material, and which was in the correct relationship to all other units, was worth one point. For ease of scoring, only errors were tallied and then subtracted from the total possible. This raw score was then converted into a "percent correct" score. Conspicuously *not* counted as errors were translated units containing awkward or nonidiomatic English, those infelicities that would have cost test-takers one point each under the current translation scoring system; nor were testees docked points for not following the GPO Style Manual. Remember that the focus here is on meaning rather than on form.

Understandably, some French language testees misunderstood the word "laïque."

(U) The stipulation that units be in the correct relationship to all other units is necessary to catch errors that might otherwise go unpenalized. Because of the redundant nature of linguistic systems, most mistakes due to syntactic misinterpretation will have effects at the lexical level and will thus be reflected in the final score. However, in rare cases it is possible to make an error of interpretation and yet represent faithfully the meaning of the original units, as when "the dog / bit / the man" is rendered as "the man / bit / the dog." In such a case, both noun phrases would be docked one point each for being in the wrong relationship to the verb.

(U) The astute reader will see that in this scoring system, unlike the PQE system, all sorts of errors receive exactly the same penalty of one point. I do not subscribe to the theory that all syntactic errors are worse than all lexical errors, and my data bear this out. For instance, one text in the experimental set had to do with a strike in Corsica which was dubbed "dead island day." Two of the lowest scorers rendered the French word for "day" ("journée") as "journey," and so failed to understand a key phrase; thus a lexical error was an important factor in a major comprehension failure. On the other hand, those who mistranslated "la crise que traverse le pays" in another text on Algeria as "the crisis that is going through the country," instead of "the crisis that the country is going through," made a syntactic error that does not seriously affect understanding.

(U) Also unlike the current scoring system, mistakes on repeated instances of the same word may be penalized more than once. This is because a word may be interpreted differently according to the immediate context in which it is found; a word may be understood in one context, but not another. Thus in one text criticizing religious schools in Paris, some testees were not able to understand the word "laïque" ('lay' or 'secular') in the first instance, when it occurred in contextual isolation, and yet recognized it in the second instance when it occurred in juxtaposition to "religious." Each occurrence of a word must be viewed as a new opportunity for understanding (or misunderstanding).

Test reliability and validity

(U) Since I did all of the scoring of the rough translations myself for practical reasons (not the ideal where judgements about correctness are involved), I should demonstrate that (1) my scoring was consistent and (2) the scores would have come out roughly the same if others had participated in the scoring.

(U) I used two measures to estimate internal consistency, neither of which is the perfect method in this situation (there are no more appropriate methods). Reliability estimates on the nine texts range from .85 to .96, averaging .90. These are quite respectable numbers. (A reliability of at least .85 is *de rigueur* in language testing; it is the minimum acceptable for the Educational Testing Service's raters of written compositions, for

example). Given that certain assumptions of the formulas were violated, reliability is probably even higher than reported.

(U) As for inter-rater reliability, I took a representative sample of eighteen translations (two for each of the nine texts, constituting about ten percent of the total used in the analysis) to independent experts, whom I briefly instructed in my scoring method. The correlation between their scores and mine was .85, once again an honorable figure. I believe it would have been even higher if the raters had had a little more "socialization," or prior discussion and practice, and if they had been allowed to discuss each other's work afterward so that they could catch their own errors and omissions.

(U) I therefore feel quite confident in saying that the test method employed in this experiment enjoys good reliability. Scoring is also fairly efficient: I completed scoring of all 201 translations in one week.

(U) Now that reliability has been demonstrated, we must ask if the test is also valid: does it really measure reading ability? One way to determine this is to see if it lines up test-takers in the same way as another reading test. The rank-order correlation between subjects' average translation scores and their combined multiple choice scores was a fairly strong .83. Thus we can say that the test's concurrent validity is good.

A model testing program

(U) The test method presented here is not the only viable alternative to the current method, but it does seem to be a good way to test reading ability, and it might easily be converted for use as a more traditional translation exam by allowing test-takers more time and being stricter in the definition of an error (this would have to be pilot-tested before implementation). In this section I would like to delineate what our overall testing program might look like, with the inclusion of a test of this type. Please keep in mind that this is only one of many possible scenarios which I hope will receive serious consideration by the pertinent decision-makers.

~~(FOUO)~~ I would like to propose the following plan for NSA's language testing:

- **Suggestion 6:**

At level:	Replace the current:	With:
2	LPT	(nothing)
3	PQE Part I, traditional translation	rough translation
3	PQE Part II, traditional translation	traditional translation (use new scoring method)

~~(FOUO)~~ I am proposing that the PQE Part I be conducted using the method described in this article. Since the PQE Part I would now be a reading test, there would be no need for a separate level 2 test. Both level 2 and level 3 reading ability could be determined simultaneously with the same test, saving a lot of time and personnel resources in test design, administration, and scoring. The people who come to the Agency with a higher foreign language capability would not have to go through three test sessions, while testees who did not achieve a level 3 rating on the first try would have to retake the test at the next offering in order to advance, as in current practice. The cutoff score for a level 3 reading ability would have to be high, at about the 90% level (remember that the mean on the level 3 texts in this study, which included many test-takers who are not language analysts and/or have not used the foreign language in many years, was already 82%). A lower cutoff score of about 75% would determine level 2 ability (this is, of course, average performance across three or more texts on which three or more experts agree about the level, etc.).

~~(FOUO)~~ The Part I test could be administered in just one hour, if texts of the same length as in this study are used. (Following tradition, rather than specifying a length in words, we could say that each passage should result in a translation of *n* words—in this case, 250 to 300.) For the Part II test, we could allow about twice as much time; a new scoring method would have to be found—perhaps an adaptation of the one described in this paper. As for the question of whether to use open or classified sources, if finding appropriate test materials in classified sources is difficult, I can think of no compelling argument for forcing PQE committees to limit themselves to that domain. The interest of ensuring fair tests should outweigh that of sticking to SIGINT.

Conclusion

~~(FOUO)~~ The text-levels theory seems to be a sound basis for our testing program as long as we take multiple samples of reading behavior at a given level, by using a variety of texts. Other aspects of our testing program are in need of some attention, particularly the scoring system used on translations. If further information or clarification is needed, I would be happy to discuss any of the ideas presented in this report. Needless to say, I believe the plans presented here would represent an improvement over the current system. However, I do not believe that we should ever be completely satisfied with what we have decided upon. Language testing is messy enough that there is always room for improvement. It is my hope that others will find the will and the cooperation necessary to carry out other large-scale studies of this sort so that at any moment we can say that NSA has done its best to make its tests as fair as they can be.

Acknowledgements

~~(FOUO)~~ It took the efforts of many, both inside the Agency and elsewhere, to make this study possible. I am especially grateful to the 67 people who suffered through my 3.5-hour testing sessions. I would further like to thank those who provided assistance in the various stages of test design and evaluation, especially

[redacted]

My gratitude also goes to everyone who provided advice and logistical help, particularly [redacted] Last but not least, I would like to acknowledge the assistance of "language levels" author Jim Child as well as my dissertation committee members, [redacted]

[redacted]

Kλ

P.L. 86-36

P.L. 86-36

SIGINT That Matters: What's the Angle?

by [redacted] E4

(U) By now, most intelligence analysts probably have heard the words "journalistic concepts" during discussions about trends in SIGINT reporting. Many of the journalistic techniques being taught in reporting courses and seminars involve changes in writing style, some subtle and some distinct. NSA has incorporated other aspects of journalism into its reporting process, such as using graphics to enhance report presentation and venturing into multi-media techniques such as video reporting, audio reporting and a myriad new dissemination methods. While these efforts have generated very positive responses from our customers, one important aspect of journalistic technique, *finding the best angle for the story*, is only now beginning to get serious attention. The angle will undoubtedly become one of the most important aspects of a SIGINT report, as Congressional budget committees again and again ask for clarification on the unique information NSA offers its customers.

(U) So, what is the angle? It is the art of presenting a topic in such a way that the reader can not only relate to that topic, but can understand the *significance* of that topic. It goes beyond simply stating the SIGINT fact, e.g. "A told B that a meeting had taken place." It helps the reader understand what went on in that meeting that was important, and why he or she should bother to read the report. To put it bluntly, *the angle is what sells the SIGINT.*

(U) Two of our most important jobs as SIGINT reporters are to **recognize** the significance of what we see, and then **explain** this in our product reports. Our customers are being bombarded by what seems to be an infinite supply of information. Unless we show them that what we have to say is important to them, our reports will simply be blown away like so much junk e-mail. We can show them by providing an angle, or focus, in our reports that emphasizes the significance of the information within the report. More than a few reporters already do this for their customers, and they receive very positive feedback. Unfortunately, some of us still believe "it came from SIGINT and nobody else has anything on it" is all the justification we need for publishing.

A Changing World

[redacted]

(U) Furthermore, I knew that since there was no real competition that might scoop me, I could rely on my customers following my every word. I did not worry much about *why* this or that happened or the deeper meaning of what might result *because* it happened. I was happy in my own little world. And I was good at it (at least everybody said so). My customers all agreed that my reports were unique and therefor[redacted] valuable, thus justifying the cost of production. If I thought about putting an "angle" on my reports, I didn't think long; if I had used one, it might well have been "because I know and you don't."

P.L. 86-36
EO 1.4.(c)

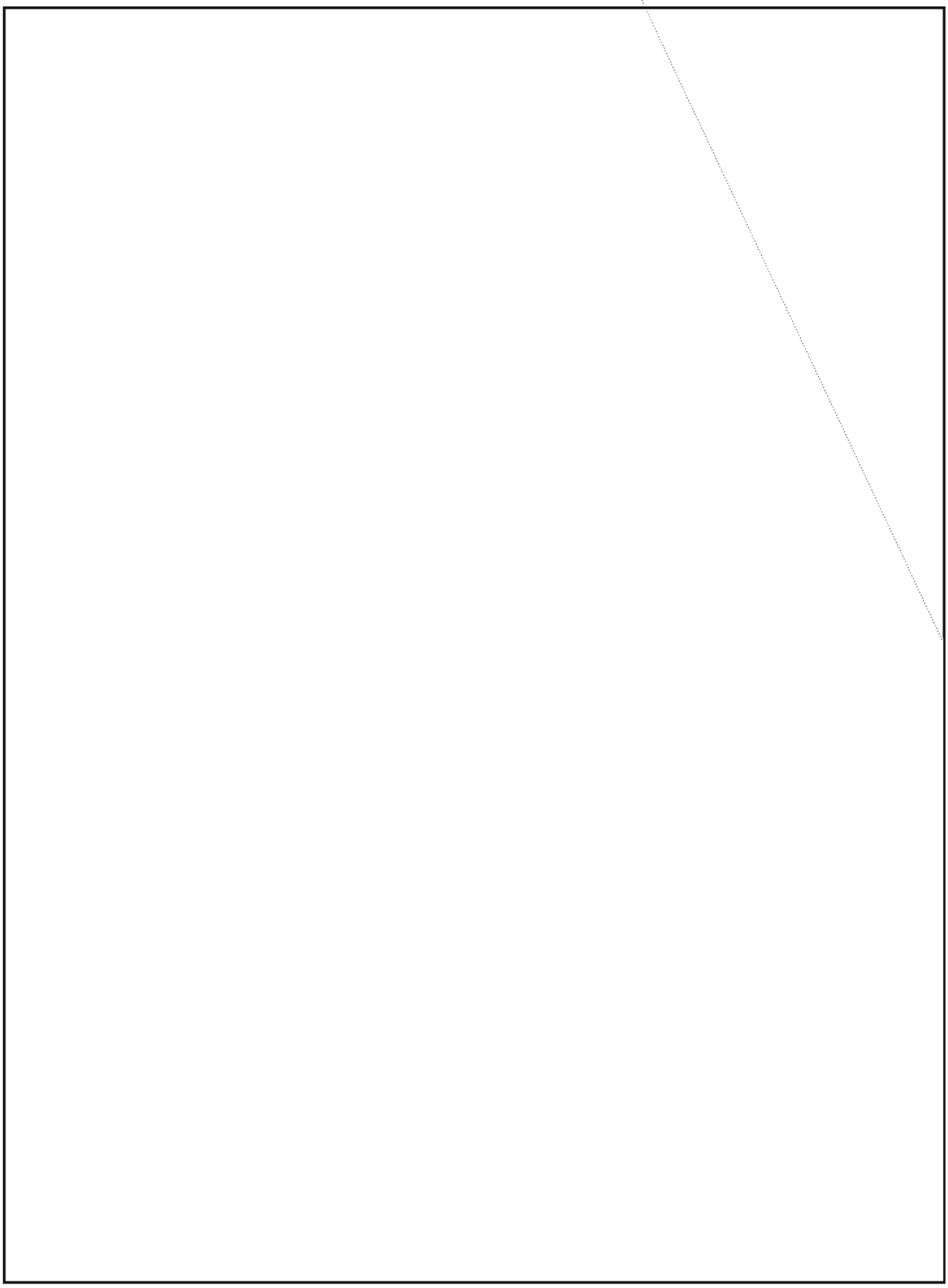
[redacted]

[redacted]

(U) If we start to suffer from blurred vision, we can close the video window and open a CompuServe window to get the latest reports from newspapers like the *Washington Post* or the *New York Times,* or from the AP or UPI wire services. Their reporters often work from an eyewitness angle, seeing things in places where even the intelligence community didn't have access just a few years ago. If we feel the need to enter the Twilight Zone, we can crank up the CD-ROM drive, slip in a disc, and cruise [redacted] to check on some minute details.

[redacted]

[redacted]

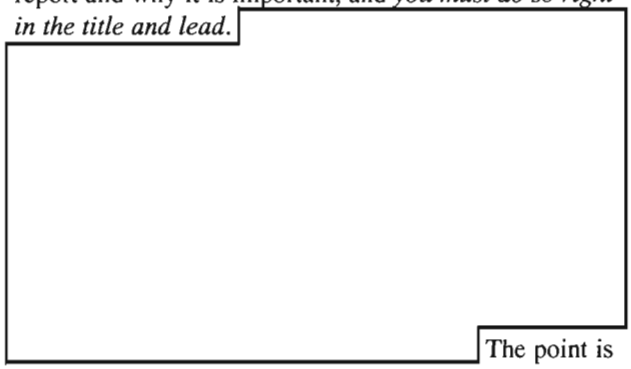


(U) As you keep pace with your target, you should also keep pace with what your customers want to know about your target. With the target, you may have the luxury of getting too much information to process. With customers, you may have the problem of having too little. Furthermore, the pace of change in what customers want may be even more rapid than it is with your target. Several new services can help you. One of the newest available is ESS topic 1442, the DO Customer Statement of Interest. It will tell you what various high-level consumers in the Washington area consider of pressing importance on any given day.

(U) As mentioned before, your customers have at least as much information to sift through as you do. They don't always have time to analyze the significance of some small fact that they may have asked you last week to report, but they will *make* time to read about that small fact if you do at least some of the analysis for them. By keeping track of what your customers might need to know, and by telling your customers what is important about each individual report, you can most effectively use the angle.

Highlighting the Angle

~~(C)~~ The final challenge is to highlight the angle for the reader. You must emphasize what is important in the report and why it is important, and *you must do so right in the title and lead.*

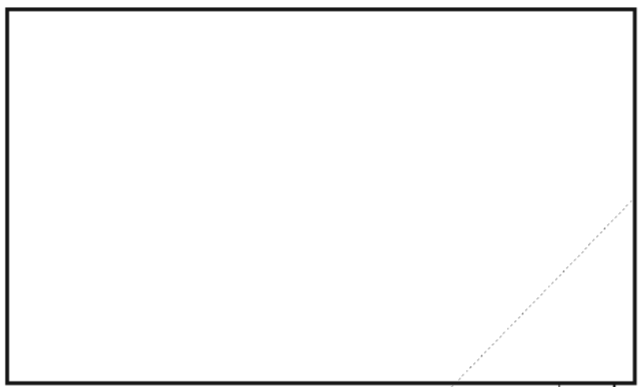


The point is to add to the greater body of knowledge about the target.

(U) *Remember that you are not obliged to make your report follow the same flow as the original traffic.* In other words, just because the Ambassador addressed the weather, the price of eggs, and the exact time that the car bomb would explode, that doesn't mean these topics must appear in the same order within your report, *or indeed that the first two must appear at all.* This is related to one of the most important lessons aspiring news reporters can learn: never leave a press conference/briefing before the question-and-answer session, even if a deadline looms. The best news angles often come from afterthoughts, side stories, or audience reactions. Similarly, the best SIGINT stories often are buried at the bottom of those long diplomatic messages or may come from some associated context. Your job is to identify them and promote them to the top of the page! For assistance in learning how to do this, P054's Reinvention Lab for SIGINT Reporting, is publishing Hints for Better Writing, available via ESS topic 1619 and MOSAIC at **http://gonzo.p.nsa/RLSR/RLSR.html**.

So What?

(U) No, Toto, we're not in Kansas any more. The world has changed. "Because I got it from SIGINT and nobody has anything on it" can no longer be the "so what?" of a product report. We must realize that we cannot report in a vacuum. We must work hard to make sure our SIGINT reports have impact. By keeping up with the target, and keeping up with what our customers want to know about the target, we will know what angle to take in our reports. And, by highlighting that angle up front in the product's title and lead, we will answer the customer's "so what" and meet their intelligence needs right off the bat. It all will add up to SIGINT worth reading.



Kλ

EO 1.4.(c)
P.L. 86-36

SIGINT and the Information Explosion

P.L. 86-36

(U) A few years ago I was attending a graduation ceremony. The guest speaker was exhorting the graduates to continue to pursue their education and self-development after graduation. To reinforce this point the speaker cited research on the growth of knowledge during the course of history. According to this individual, human knowledge doubled between the beginning of recorded history and about 1900. Knowledge doubled again between 1900 and the end of World War I. It doubled again between that time and about 1930; then again by 1940, and since then has been doubling every three to five years.

~~(C)~~ Now what, you might say, does this have to do with SIGINT, or, more specifically, with Cryptologic Support Groups (CSGs)? Many things, but as the chief of a CSG myself, I see the greatest challenge in this information explosion is the growth of sources of information beyond our wildest dreams. In the days of the Cold War CSGs had a much simpler job than we do today. We interpreted SIGINT. We had a good time and we made many friends for the Agency. We worked with "all-source" analyst, but I submit to you that all-source analysis in those days was much more straightforward than it is today.

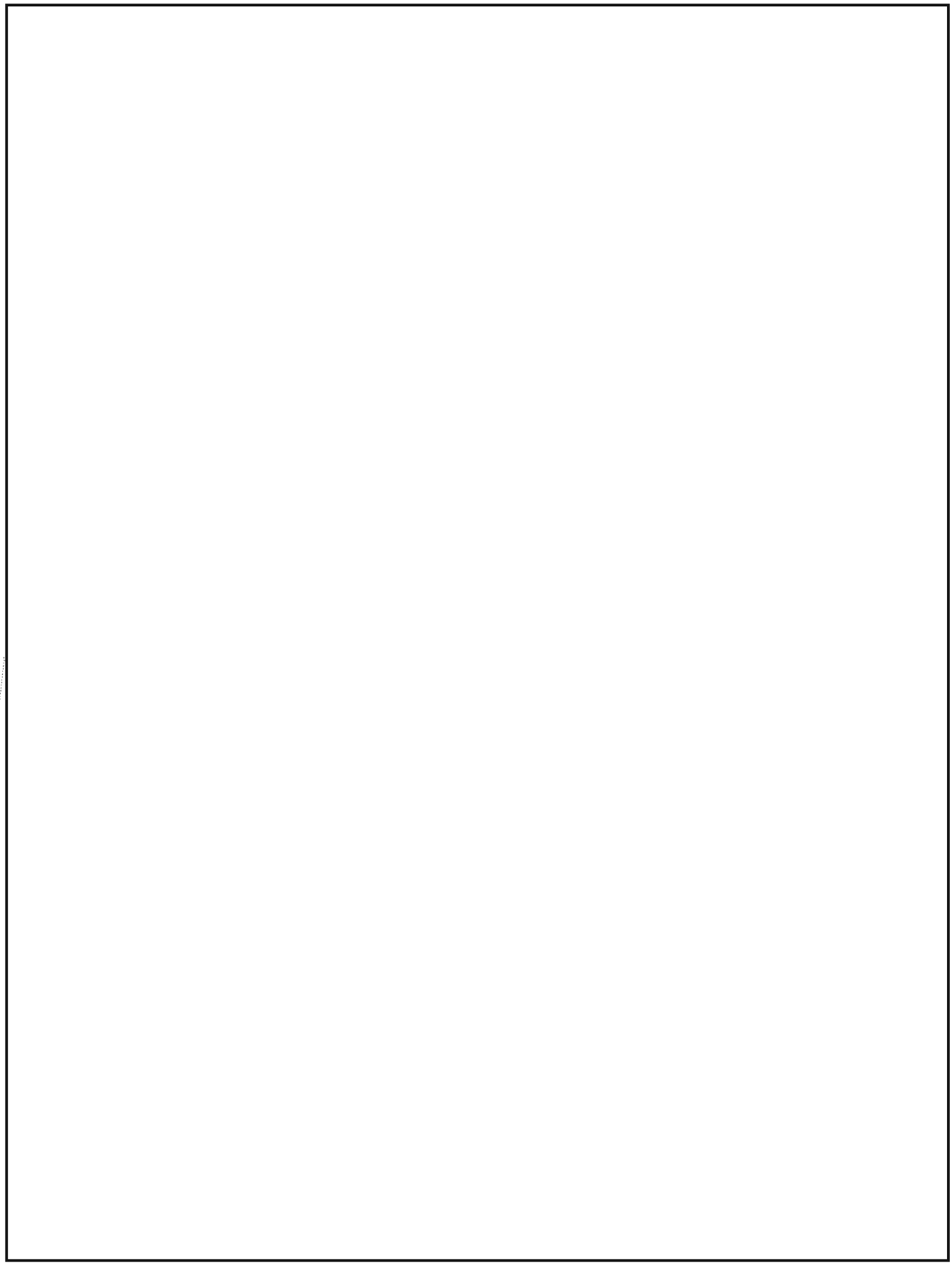
Must call George before we attack. Anybody see a pay phone?

(U) First: the analysis described above is straightforward. Each collection system provides a time, a location and the description of an activity. Analysis mainly consists of looking at the collectors' outputs and verifying that they really do correspond.

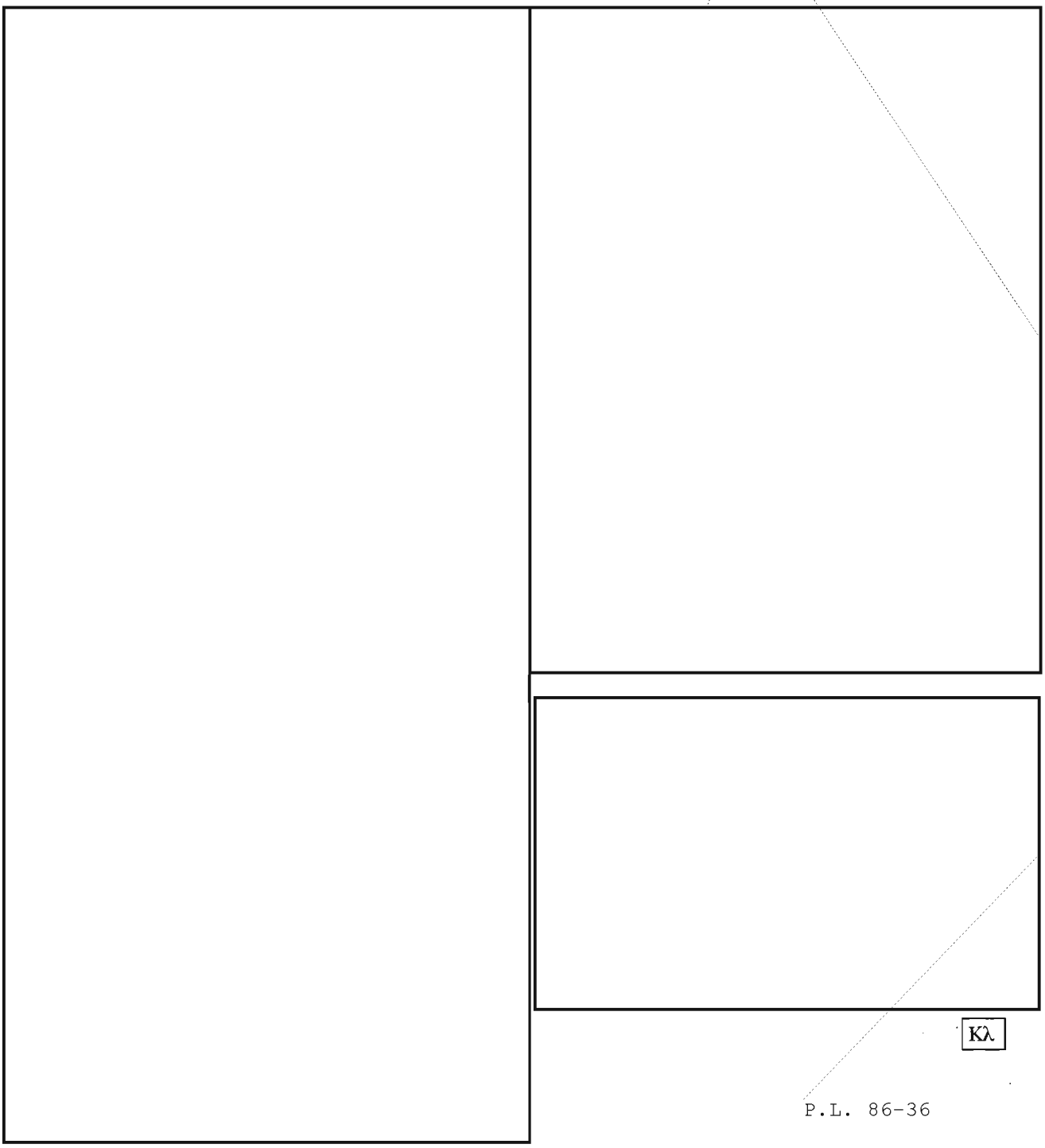
(U) Second: the sources described above are complementary, that is, they fit together neatly to form a coherent piece of "all-source" data.

EO 1.4.(c)
P.L. 86-36

EO 1.4.(c)
P.L. 86-36



P.L. 86-36
EO 1.4.(c)



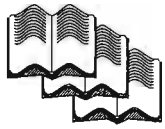
ΚΛ

P.L. 86-36

P.L. 86-36

Recent Publications on Information Warfare

reviewed by [redacted]



(U) Information Warfare or IW is very topical these days. One cannot pick up a newspaper or magazine without seeing an article that at least touches on IW-related issues. Even the Speaker of the House of Representatives regularly holds forth on the subject. It is not surprising that the publishing industry has picked up on the IW trend and printed numerous books on IW, fiction and non-fiction. (Admittedly, it is often hard to tell the difference.)

(U) One of the first problems one encounters in discussing IW is its definition. There is no widely accepted taxonomy of IW. Consequently, it covers a multitude of high-tech material including command-and control-warfare, perception management, computer warfare, gathering intelligence from computers, using computer viruses to destroy data, affecting an adversary's infrastructure through the use of computers, esoteric weapons such as electromagnetic pulse devices and microwave beam guns and so forth. Of course, IW must also address how we protect our information and information systems from the depredations of others.

(U) The basis for much in IW stems from the world's growing interconnectedness. It is this exponentially expanding network of networks that leads many futurists to predict fundamental changes in the way the world works and even in the nature of what constitutes national interests. Heidi and Alvin Toffler have written much on this subject. If they are right and information is the foundation of the next stage of human development, then our understanding of IW-related issues may determine the nature of our future as a world power.

(U) IW is a sprawling field of wide-ranging ideas. It, and its companion concept cyberspace, make up the new Wild West. Fortunes will be made and lost, power centers will pivot and shift, indeed lives will be shaped by how well we adapt to a future based on information. Below are some comments on a few of the recent offerings in IW from the publishing world.

(U) *Information Warfare: Chaos on the Information Superhighway* by Winn Schwartau. Definitely worth reading. Nevertheless, keep in mind that Mr. Schwartau is something of an IW gadfly who does not always have his facts straight. Many NSA readers of this book will spot errors which cannot be discussed in this review. In addition to factual errors, he is inconsistent in his approach to IW. For example, in *Information Warfare* he decries the fact that we as a nation are essentially unprepared for even low-level attacks by hackers and criminals. On the other hand, he recently led an e-mail campaign suggesting people flood the government's e-mail addresses and bring the system to a halt in protest of the very cryptographic policies which might offer a modicum of the protection he says we need. Now, having relieved myself of that editorial comment, let me talk about the book itself.

(U) *Information Warfare* frames many of the IW issues clearly and with imaginative scenarios describing how bad things may become. Mr. Schwartau divides those who would do harm to information systems into several categories or levels. He differentiates between the curious hacker, the criminal hacker for hire, the disgruntled employee, the cyber-terrorist group, and the nation-state. He points out that a well-focused nation-state or terrorist group will be able to do more harm with computer attacks against power grids or the banking system than they would ever achieve with explosives. He also points out the vulnerabilities we all face by having our individual credit histories, medical records and other personal data so easily available in large data bases. The fact that these data bases can be accessed and even altered is no secret. Anyone who has tried to rectify a credit bureau error understands the potential nightmare an attack aimed at one person might produce. Mr. Schwartau is particularly evocative in describing this type of scenario.

(U) While Mr. Schwartau performs a service by pointing to the problem, he fails to offer meaningful recommendations to fix it. He offers some vague advice about government action, but there is nothing concrete. One is left with the impression that he wants a government he doesn't trust to go after "them" and to do so without infringing on anyone's personal privacy in the process. Bottom line: a little shrill but still a good read and a good introduction to some of the fundamental issues.

(U) *War and Anti-War* by Heidi and Alvin Toffler.[1] Don't get me wrong—there are some important ideas here, it's just tough wading through a couple of hundred pages of disjointed prose to dig them out. Instead of reading this book, you may want to read *Creating a New Civilization, Politics of the Third Wave* with the introduction by Newt Gingrich. It is the condensed version not only of *War and Anti-War* but also of *The Third Wave* and other works by the Tofflers. The Tofflers are idea people and they have some good ones, even if they are given to sweeping generalization. They just need a better editor.

(U) If you can tolerate *War and Anti-War*'s "thought-bite" format aimed at those with short attention spans, you will gain some useful insights. For example, the Tofflers believe the way we make war is based on the way we create wealth. In ancient times, wealth was based on agriculture and war was centered on agricultural concerns. In the industrial age it was technology that drove the economic engine and likewise the engines of war. Now as we enter the information age, how well we degrade an adversary's information and information systems while protecting our own will determine our survival.

(U) The Tofflers have derived this concept from some of their earlier thinking about second wave (industrial) and third wave (information) civilizations. One of the problems the Tofflers foresee is that what they perceive as second wave organizations (like the federal government, or NSA, for that matter) are not well suited to deal with third wave (IW and cyberspace) problems. They predict that the first nations to adapt to the information age and the third wave will be the superpowers of the next century. I recently had the opportunity to ask Alvin Toffler what an effective third wave government might look like. He said he had no idea. I guess that's a reasonable answer since one has yet to emerge.

(U) The Tofflers see most modern conflict as arising from the clash between waves. For example, the mess in the Balkans is seen as first wave (country) vs second wave (cities). The American Civil War would also be a first/second wave confrontation. The Tofflers depict many of the current problems in contempory American politics as second/third wave conflicts. They are not sanguine about chances for a peaceful transition to a third-wave world. They point out that such transitions have traditionally been marked by considerable chaos and upheaval. So, if you can get through the chaos and upheaval of the book's poor organization, read it.

1. See "Information Warfare, *War and Anti-War*, and NSA", by Bobby Mitchell, CRYPTOLOG Fall/Winter 1994.

(U) *The Cuckoo's Egg* by Clifford Stoll. Still a classic. This is the story of how the Hannover Hackers were caught and their KGB connections exposed. Reading this book will give you a good understanding of what it takes to detect and then track down a hacker and beat him at his own game. It will also give you an appreciation of the difficulties involved in trying to protect ourselves while staying plugged into the rest of the world. It reads like a detective story and imparts solid knowledge. (Incidentally, neither the computer-security folks at the installations the hackers were invading nor various U.S. government agencies come off too well.)

(U) *Being Digital* by Nicholas Negroponte. An engaging, if shallow, ramble along some of the implications of the brave new world of interconnectedness. If you haven't thought much about the subject, this is a good book. If you have thought about it at all, wait for the library copy and skim it.

(U) *Silicon Snake Oil* by Clifford Stoll. The author of *The Cuckoo's Egg* waxes eloquent about what we seem to be losing in all the hype about cyberspace and being interconnected. He stops short of coming down on the side of the luddites who would have us all unplug and return to the age of Dickens, but he does decry the loss of face-to-face human contact. He makes a reasoned case that much of the cyberspace story has been oversold and that we are in for some disappointments. He says the computer will never provide a real sense of community and in fact works against it. He also reminds us that limiting our connectedness or choosing not to connect at all remain viable alternatives.

(U) *Masters of Deception* by Michelle Slatalla. This is an entertaining and disquieting look into the mind of cyberspace gang members. It centers around the story of a now-famous hacker who is doing time in the "big house" for his exploits. The book is well written and the tale moves along like good fiction. The ethos of the hacker comes through well, as does a solid feel for the hacker culture.

(U) In the fiction category the following IW-related titles have recently been published.

Debt of Honor by Tom Clancy. The Stock Exchange can be used to make a point as well as money.*Black Cipher* by Payne Harrison. An evil cabal in GCHQ tries to hide from a brilliant cryptanalyst (not to mention a

good card player).

Interrupt by Toni Dwiggins. Who says the phone system is always there when you need it?

Heavy Weather by Bruce Sterling. Mad Max meets the cyberpunks in an almost-doomed post-eco-disaster world.

Neuromancer by William Gibson. Still the classic of virtual reality for those who would rather have a cyberlife than no life at all.

[redacted]

P.L. 86-36

P.L. 86-36
EO 1.4.(c)

Book Review

China's Air Force Enters the 21st Century. by Kenneth W. Allen, Glenn Krumel, and Jonathan D. Pollack. Santa Monica, California: RAND, 1995.

P.L. 86-36

Reviewed by [redacted]

~~(C CCO)~~ The principal author of this groundbreaking study of the Chinese Air Force is Kenneth W. Allen (a retired U.S. Air Force major), [redacted] Mary McGarrahan, [redacted]

~~(C)~~ Ken Allen later served with distinction as an intelligence officer with PACAF and at the Defense Intelligence Agency. As a Captain, he was an assistant Air Force attaché at the U.S. Embassy in Beijing at the time of the pro-democracy movement and the demonstrations in Tiananmen Square. This volume is an outgrowth of research begun in Beijing by Mr. Allen in the latter part of the 1980s.

(U) Simply put, ***China's Air Force Enters the 21st Century*** is the single best book on the Chinese Air Force which has yet appeared in the English language. Allen and his co-authors have made unprecedented use of original source material not heretofore available to the general public. They present a cogent and conservative analysis of the Chinese People's Liberation Army Air Force (PLAAF), its history, its organization, and its potential. Unlike a number of commentators, the authors do not treat the Chinese armed forces as if they were marching inexorably toward the domination of East Asia. Messrs. Allen, Krumel, and Pollack quite rightly point to several critical problems facing the PLAAF in the next several years in the areas of leadership, manpower, technology, budget, and competition.

The PLAAF plans to replace obsolete aircraft with models like the upgraded Super-7

(U) Among the specific challenges facing Beijing, the authors cite block obsolescence of aircraft types. In the foreseeable future, the PLAAF will have to replace the F-6/FARMER fighter, a Chinese version of the 1950s era Soviet MiG-19 design. Currently, the F-6 fleet comprises some 65% of the total fighter inventory. The PLAAF is trying to address this need by means of indigenous programs incorporating foreign technology (upgrades to the F-7/FISHBED and the F-8/FINBACK, the Super-7, the FB-7) and through outright purchases of foreign aircraft (notably, the Russian Su-27/FLANKER). Incorporating these new aircraft types into PLAAF tactical operations, however, will require significant changes in procedures and a great deal of training.

(U) Allen and his colleagues make clear that the Chinese Air Force is operating under serious financial constraints. As they put the matter, "It is difficult to understate the scale of resources that would be required for the air force to make an effective transition to a credible, modernized force structure." Simply buying Russian or European or even American systems "off the shelf" may be relatively more cost-effective in the short run than designing and producing these systems indigenously. In the long run such a policy would be extremely damaging to the health of the Chinese aviation and aerospace industries, and would undermine any attempts by the PLAAF to become self-sufficient. A combination of indigenous production and imported technology appears to be the preferred Chinese solution.

(U) If the volume has a measurable flaw, then it is one of omission, not commission. The book cries out for photographs of PLAAF aircraft, senior officers, and the like. Nonetheless, the individual chapters and appendices are chock-full of valuable detail on leadership, force structure, strategy, education and training, budget, the political commissar system, PLAAF ranks, aircraft procurement programs, fighter aircraft, and air defense.

(U) The authors state that the PLAAF does not pose a serious threat to the United States or its interests at present. They point out that this relative situation will not change dramatically over the next ten years or so. In the longer run, however, China may well develop a potent air force—if the PLAAF continues modernizing its aircraft, weapons systems, force structure, aerospace industry, and doctrine. Much will depend on the political will of the leadership in Beijing and their allocation of resources.

~~(S)~~ The so-called "RAND study" apparently has become somewhat of a *cause célèbre* for the senior leadership of the Chinese armed forces. According to the U.S. Defense Attaché Office in Beijing, the Chinese Central Military Commission has begun an investigation into "how such sensitive information on the PLAAF could have gotten into the hands of the RAND people." Allen's book is being translated into Chinese in order to facilitate this investigation. Furthermore, the PLAAF is taking another look at its budgetary requests in light of the critical comments made by Allen and his colleagues.

[redacted]

(U) Mr. Allen and his colleagues could not have asked for a better review of their ground-breaking work on the Chinese Air Force than that. A notable book, indeed.

(N.B. Mr. Allen would be happy to receive questions or comments on his book. Please send your remarks to **cryplog@nsa**. We will pass them on to the author.)

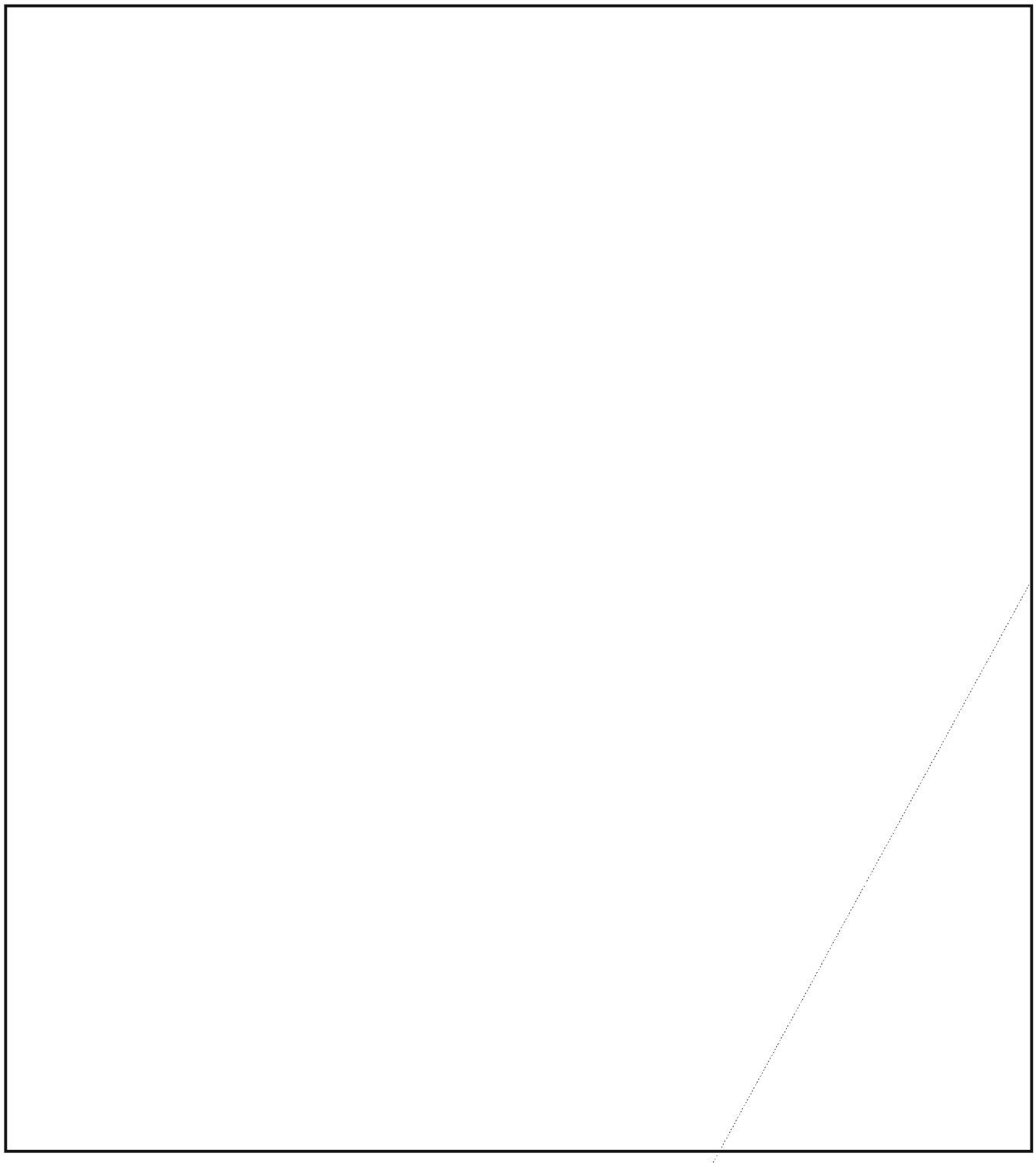
[redacted]

Kλ

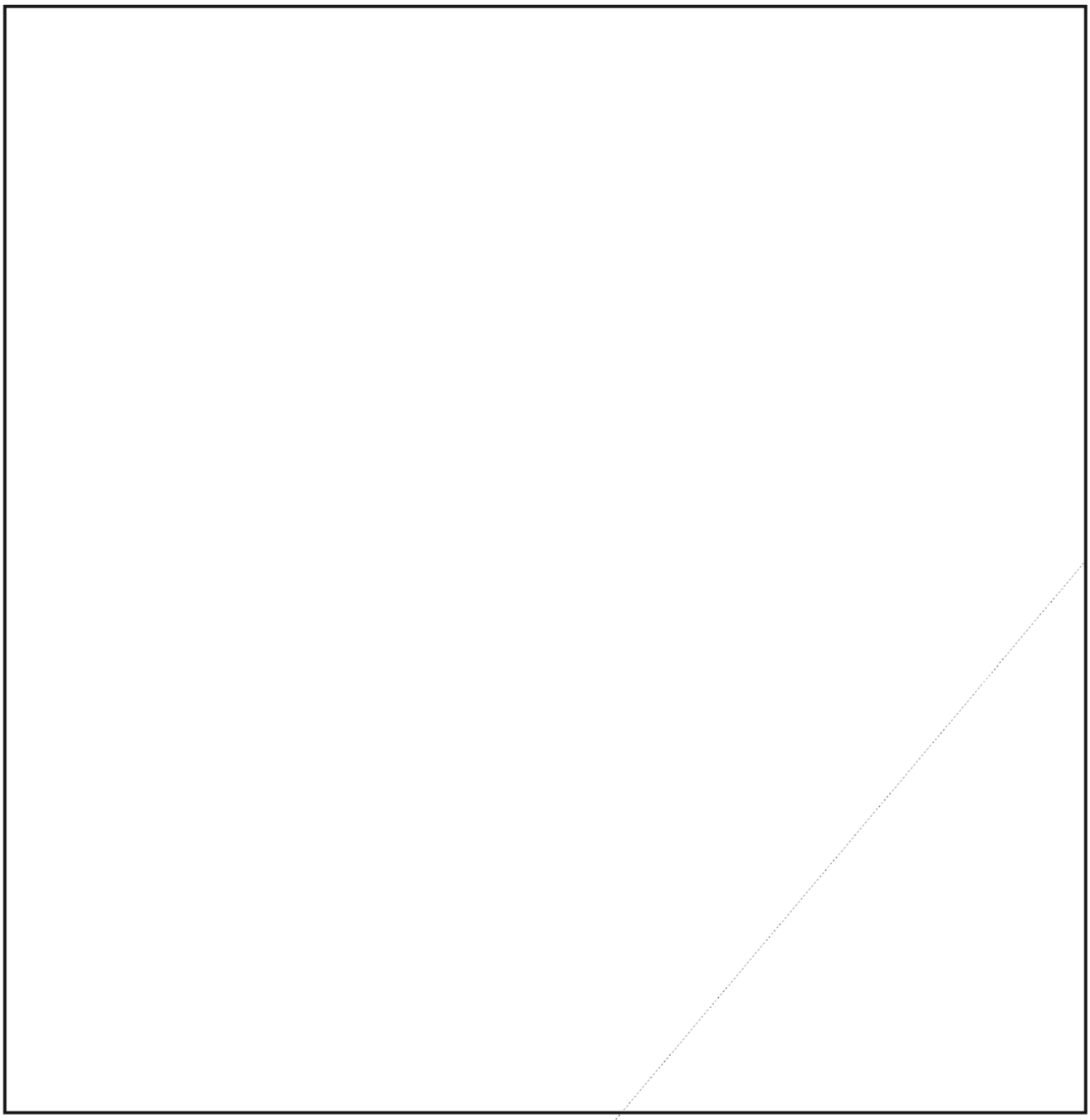
P.L. 86-36
EO 1.4.(c)

Cryptologic Lessons Learned

An Excerpt from N25's Data Base



EO 1.4.(c)
P.L. 86-36



Kλ

EO 1.4.(c)
P.L. 86-36

SIGINT Bloopers

We begin, as before, with the Homonym Pitfall:

[redacted] Does the State Department know how frequently the international affairs of that town in Howard County show up in our product? Imagine the USSID 18 restrictions!

Either the reporter or the editing staff must have been hungry when they sent out a product that referred to the capital of India as "New Deli."

"The rebel leader and his principle ally"—nice to know the ethical aspects of rebellion are being handled properly.

Again, it's not just product reports: an official security regulation warns against drawing "undo" attention to one's place of employment, while an office-level action memo asked the organization concerned to "pole" its people.

In the Mark Twain (Almost But Not Quite Right Word) category:

"The UN representative said the political solution was under control but worrisome"—you've got to watch out for those solutions; they can get out of hand.

"Reports on Withdrawal of Heavy Weapons Seem Contrary"—maybe those reports got up on the wrong side of the bed.

One product report referred to "the three phases of the process, vice disarmament, demobilization, and reintegration of the combatants." We racked our brains trying to figure out what those phases could be, if not disarmament, etc.? A phone call to the OPI revealed that the reporter had been thinking of "viz." vice "vice." (On a serious note, this is a prime example of why we publish this column: such errors can result in a report saying the opposite of what was intended, as in this case.)

Other howlers can only be classified as Just Plain Weird Writing. "Bovine medical assistance to Fredonia is imminent"—what Gary Larson-like images this one conjured up! Apparently the reference was to the expected arrival of a team of foreign veterinarians, but references to "udder devastation" and "milking the situation" floated around the office for some time. Our favorite comment on this one was that it gave "Medecins Sans Frontieres" a whole new meaning: Doctors Without Fences...

In another case, the giver and taker of a bribe were described as involved in "bilateral corruption" (as opposed to the ordinary kind).

Anthropomorphism is evident in some OPIs:

"The controlling authority ordered the relay station not to harm the hostages." Who knows what that radio equipment will get up to if you don't keep an eye on it.

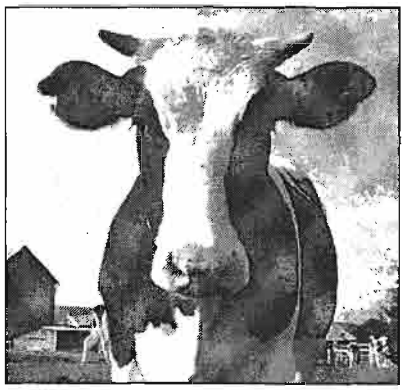
"The nuclear briefcase accompanied the president and was active during the visit."—bet that briefcase was the life of the party.

"The ambassador's Zendian residence learned that he would be leaving." He must have one of those "smart houses" that he can tell when to lower the heat and turn on the voice mail.

From the Department of Redundancy Department:

"Although the situation in Fredonia was untenable due to anarchy on the streets of the capital, Fredonia could remain in untenable straits for quite some time longer." (This was issued by an OPI that thinks "undercurrent" is two words.)

From the same report: "The extent to which the prime minister will act is unknown at this time and no evidence of such is currently available." "He may become a future rival"—well, as long as he stays a future rival, there's no need to worry.



Drink a glass of milk before bed and call me in the morning

As before, thanks to all contributors; examples may be sent to P054 in Rm. 3E027, Ops. 1, or via e-mail to **cryplog@p.nsa**.

Kλ

Editorial Policy:

(U) Technical articles are preferred over those relating to management, shorter over longer (under 3,500 words). Emphasis should be on improving NSA's technical performance; articles should be aimed at explaining developments in one's career field to thos outside it. Readers are invited to contribute conference reports and reviews of books, articles, software, and hardware that relate to our missions or to any of our disciplines. Editorials are also welcome, as is humor. Submissions may be published anonymously, but the identitiy of the author must be known to the editor.

Submitting Articles:

(N.B. If the following instructions are a mystery to you and your local ADP support is no help, please feel free to contact the CRYPTOLOG editor on 963-3123s or cryplog@p.nsa.)

~~(FOUO)~~ Send a hard copy accompanied by a labelled diskette to the editor at P054 in 3E027, Ops. 1, or send a soft copy via e-mail to **cryplog@p.nsa**.

Guidance:

For maximum efficiency (as far as possible within the limits of your word processor):

- Do not type your article in capital letters.
- Classify all paragraphs
- Label all diskettes, identifying hardware (operating system: DOS, UNIX), density and type of word processor used, your name, organization, building, and phone number.
- FrameMaker format is preferred; ASCII text is also fine. J334 has a conversion service that converts Interleaf, WordPerfect, OfficeWriter, and MS Word into FrameMaker. Just attach the document to an E-Mail Compose Window addressed to **convert@nsa**.