DOCUMENT RESUME

ED 358 110 TM 019 876

AUTHOR Mann, Doug
TITLE The Relationship between Diagnostic Accuracy and Confidence in Medical Students.
PUB DATE Apr 93
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classification; *Clinical Diagnosis; Comparative Analysis; *Confidence Testing; *Decision Making; Higher Education; Medical Education; Medical Evaluation; *Medical Students; *Metacognition; Osteopathy; Physicians; Probability; Psychology; *Self Esteem
IDENTIFIERS *Accuracy; Adjusted Resolution; Rater Reliability

ABSTRACT

Studies in psychology and clinical decision making have shown that research subjects and physicians are often overconfident in the accuracy of their judgments. In these studies, groups of 20 first-year and 27 third-year osteopathic medical students at the Ohio University College of Osteopathic Medicine (Athens) were slightly underconfident in their ability to classify artificially-generated abnormal heart rhythms, with good "calibration" at higher confidence levels. Evaluating individuals on "adjusted resolution," which provides an index of the ability to sort correct and incorrect diagnoses into different confidence levels, reveals a curvilinear relationship between diagnostic metacognition (adjusted resolution) and accuracy. The best metacognition was found in subjects achieving 70 to 85 percent diagnostic accuracy. Implications of these findings are discussed. Three figures illustrate the discussion, and an appendix contains a diagram of the decomposition of mean probability scores. (SLD)

# The Relationship between Diagnostic Accuracy and Confidence in Medical Students

Doug Mann

Ohio University College of Osteopathic Medicine

Presented at the Annual Meeting of the

American Educational Research Association

session 6.17: Studies in Professional Expertise

Henk G. Schmidt, Chair

Atlanta, April 1993

Doug Mann
Educational Development & Resources
OU-COM
225 Grosvenor Hall
Athens, Ohio 45701

ph. 614/593-2229
fax 614/593-9180
internet: MANND@OUVAXA.CATS.OHIOU.EDU

## ABSTRACT

Studies in psychology and clinical decision-making have shown that research subjects and physicians are often overconfident in the accuracy of their judgments. However, in these studies, groups of first and third-year osteopathic medical students were slightly underconfident in their ability to classify abnormal heart rhythms, with good "calibration" at higher confidence levels. Evaluating individuals on "adjusted resolution," which provides an index of the ability to sort correct and incorrect diagnoses into different confidence levels, revealed a curvilinear relationship between diagnostic metacognition (adjusted resolution) and accuracy; the best metacognition was found in subjects achieving 70-85% diagnostic accuracy. Implications of these findings are discussed.

# The Relationship between Diagnostic Accuracy and Confidence in Medical Students

Because medical diagnosis is probabilistic, appropriateness of confidence in a diagnostic hypothesis is critical. Underconfidence in a correct diagnosis leads a physician to seek additional information to confirm the diagnosis and rule out alternatives, wasting considerable time, effort, and money, and possibly allowing the patient's condition to worsen. Overconfidence in an incorrect diagnosis could be disastrous, causing a physician to disregard available relevant information and proceed with an inappropriate treatment decision.

Most confidence research in psychology has required subjects to indicate confidence on a 50-100% scale after answering each of a large number of two-alternative general-knowledge questions, e.g., "Anna Freud is Sigmund Freud's (a) oldest child, (b) youngest child" (Lichtenstein & Fischhoff, 1977). In general, subjects' mean confidence on these tasks exceeds their mean accuracy, which is referred to as overconfidence (Lichtenstein, Fischhoff, & Phillips, 1982).

Subjects' confidence judgments can be grouped into intervals, e.g., .5-.59, .6-.69 and so on; the mean percentage of correct responses across subjects at each confidence interval usually does not match the interval. In other words, groups of subjects are under- or overconfident at most intervals, and thus are said to have poor "calibration" (Lichtenstein & Fischhoff, 1977). Yates (1982) refers to the relationship of confidence intervals and accuracy as "calibration-in-the-small," and the issue of global (mean) over- or underconfidence as "calibration-in-the-large."

Confidence or subjectivity probability data can be displayed using calibration curves, covariance graphs or receiver operating characteristic (ROC) curves (Poses, Cebul, & Centor, 1988). Statistical analysis of confidence data is often based on the mean probability score ($\overline{PS}$), or Brier score (Brier, 1950). The "Murphy resolution," part of a widely-used decomposition of $\overline{PS}$ (Murphy, 1973), indexes the ability of each subject to sort correct and incorrect diagnoses into different confidence levels (Yates, 1982; see Appendix for details of $\overline{PS}$ decomposition). Yates observes that

> while the appeal of calibration-in-the-small is largely "aesthetic," the practical significance of resolution can potentially be much greater. In essence, resolution pertains to a much more fundamental skill than calibration; it refers to the ability of the forecaster to discriminate individual occasions on which the event of interest will and will not take place. By contrast, calibration concerns the forecaster's ability to assign the "right" numerical labels to his or her forecasts.

Clearly, resolution is critical to medical diagnosis. It would be of little comfort to a patient to know that a physician is correct on 70% of the diagnoses in which he or she expressed 70% confidence (which would contribute to a good calibration score). The physician must strive to know whether each and every diagnosis is correct or incorrect, and this ability is measured by the resolution score.

The resolution score can be divided by the variance of the proportion correct for each subject to provide an index of metacognition that is comparable across subjects with different levels of diagnostic accuracy; the score is then equivalent to the effect-size measure eta$^2$ (Sharp, Cutler & Penrod, 1988). This score will be referred to as "adjusted resolution" and used as the measure of diagnostic metacognition. Adjusted resolution scores range from 0-1, and higher scores reflect more consistent assignment of different confidence levels to correct versus incorrect diagnoses.

Lichtenstein and Fischhoff (1977) found a strong curvilinear relationship between calibration and mean accuracy in an experiment using general-knowledge questions, with the best calibration found in subjects who achieved moderate (approximately 80%) accuracy; subjects at lower and higher levels of accuracy had worse calibration scores. Lichtenstein and Fischhoff concluded that "Those who know more do *not* generally know more about how much they know." Replication of this curvilinear relationship in another domain with the more fundamental skill measured by adjusted resolution would be important to the study of decision-making.

Most studies of physicians' subjective probabilities and diagnostic accuracy have required physicians to encounter a series of patients or brief written cases and then estimate the probability of a single target disease being present in each "patient." In one study, physicians examined 1,531 first-time patients with coughs and overpredicted the incidence of pneumonia at every subjective probability level above

"0" (Christensen-Szalanski & Bushyhead, 1981). This is a common but not universal finding in studies of physicians' subjective probability estimates (reviewed in Yates, 1990).

Studies of this type do not provide an evaluation of general diagnostic ability, because the physicians are not required to construct a differential diagnosis based on their own interpretation of the case. A more realistic task would at least provide a list of diagnostic options rather than specifying the disease to be considered.

Few studies of diagnostic confidence have been done with medical students, although several studies have investigated the use of confidence-weighted multiple-choice test items. In one study, second-year medical students indicated confidence for each item on an Introduction to Clinical Medicine course exam; the addition of confidence-weighted scoring improved the contribution of the test scores to the prediction of National Board of Medical Examiners Part I and II test scores in a multiple regression equation, and overconfidence was related to junior clerkship test scores and ratings (Zeleznik, Hojat, Goepp, Amadio, Kowlessar & Borenstein, 1988). Indices of overconfidence and underconfidence derived from a confidence-weighted test of general pediatric knowledge taken by pediatric residents correlated with faculty members' perceptions of resident confidence, and residents with extreme overconfidence or underconfidence scores were more likely to leave the residency program (Hausman, Weiss, Lawrence, & Zeleznik, 1990). These studies suggest that the relationship between confidence and accuracy may affect the acquisition and application of clinical knowledge and decision-making skills.

Medical student diagnostic confidence is of general interest in terms of novice-to-expert development, and raises an interesting question in relation to the common finding in the medical decision-making literature that experienced physicians are overconfident in their diagnostic ability: Is the genesis of diagnostic overconfidence to be found in the medical school years, or does it develop only with years of experience in practice?

The following studies therefore explore two independent questions: Are medical students overconfident in their diagnostic ability? Secondly, is there a relationship between diagnostic metacognition (adjusted resolution) and diagnostic accuracy, and,

if so, does that relationship take the curvilinear form found by Lichtenstein and Fischoff?

## Subjects

Both studies used medical student volunteers from the Ohio University College of Osteopathic Medicine. For the study with first-year students (Mann, 1989), 20 volunteers from a class of 100 participated; four subjects with extensive prior experience in the subject matter were allowed to participate, but their data was not included in the analysis. In the study with third-year students (Mann & Dane, 1991), 27 volunteers from a class of 85 participated. The students were told that they would receive valuable practice in the classification of cardiac rhythm disturbances, and they received no compensation or course credit for their participation. The third-year students faced a practical exam on the topic of cardiac dysrhythmias (abnormal heart rhythms) in an Emergency Medicine course two weeks after the experiment.

## Materials

The classification of cardiac dysrhythmias is one of the first diagnostic skills that medical students are expected to learn. There are approximately 14 types of dysrhythmias (minor variations in the taxonomy produce 13-16 categories). An orderly analysis of five characteristics of a rhythm (rate, regularity of rhythm, P-wave, QRS complex, and the P-QRS relationship) will provide a definitive classification of its abnormality into such categories as atrial tachycardia, ventricular fibrillation or first-degree AV (atrioventricular) block.

Examples of the normal sinus rhythm and the dysrhythmias were generated by connecting a Laerdal HeartSim 2000 to a LifePak defibrillator/cardiac monitor. The examples were printed, and 35mm slides were taken of a section of each "rhythm strip" that included 8.5 seconds of the rhythm. The rhythms were judged to be very realistic by a physician currently practicing emergency medicine.

## Procedure

Both studies required medical students to classify each slide and to indicate a level of confidence in each diagnosis on an 11-point scale from 0-100%. Mann (1989) tested 16 first-year students using 27 ECG slides; Mann and Dane (1991) tested 27 third-year students using 32 ECG slides with accompanying case scenarios. Each test item was

timed: the first-year students had one minute to classify each slide, and the third-year students were allowed 30 seconds per rhythm slide. Each rhythm slide was followed by a black slide which was displayed for 15 seconds to allow students to record their diagnoses and confidence levels.

## Results

"Calibration curves" are formed by combining subjects' responses at each confidence level on the x-axis plotted against the proportion correct on the y-axis. For perfectly-calibrated groups of subjects, all points would fall on the diagonal line. Figure 1 illustrates the calibration of all subjects in each of the two experiments; the numbers on each graph are the numbers of responses represented by each point on the line.

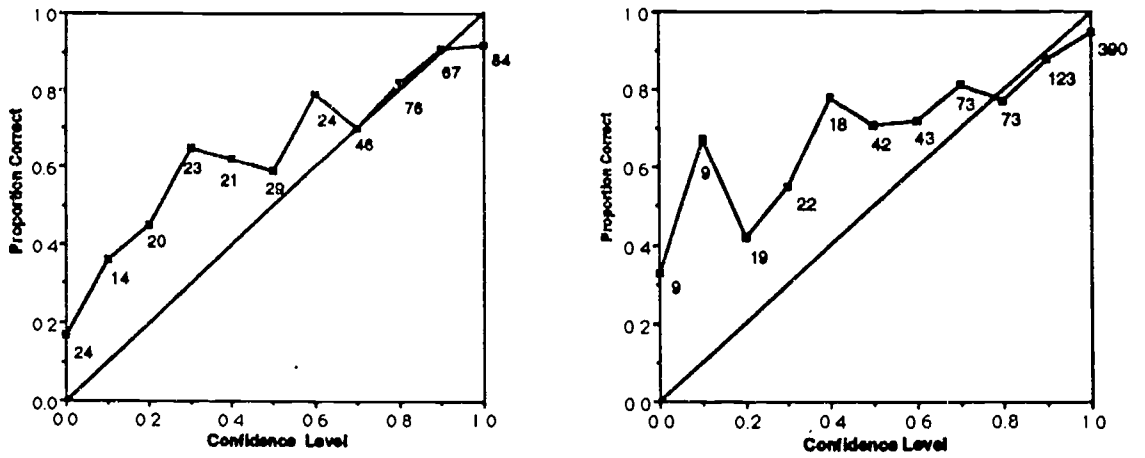


*Figure 1. Calibration curves for 16 first-year (left) and 27 third-year (right) students .*

Both graphs illustrate marked underconfidence at lower confidence levels and slight overconfidence at high levels of confidence. For the first-year students, mean accuracy was 73.4% but mean confidence was only 67.0%. The third-year students achieved a mean accuracy of 85.1% and had a mean level of confidence of 82.3%. Both groups of subjects were slightly underconfident in their diagnostic ability. These differences between diagnostic accuracy and confidence were not subjected to tests of statistical significance for two reasons: no test is necessary to establish that the students were not generally overconfident, and further analysis of calibration was omitted in favor of the analysis of adjusted resolution.

The concept of resolution is illustrated by the calibration curves of two first-year

subjects in Figure 2.

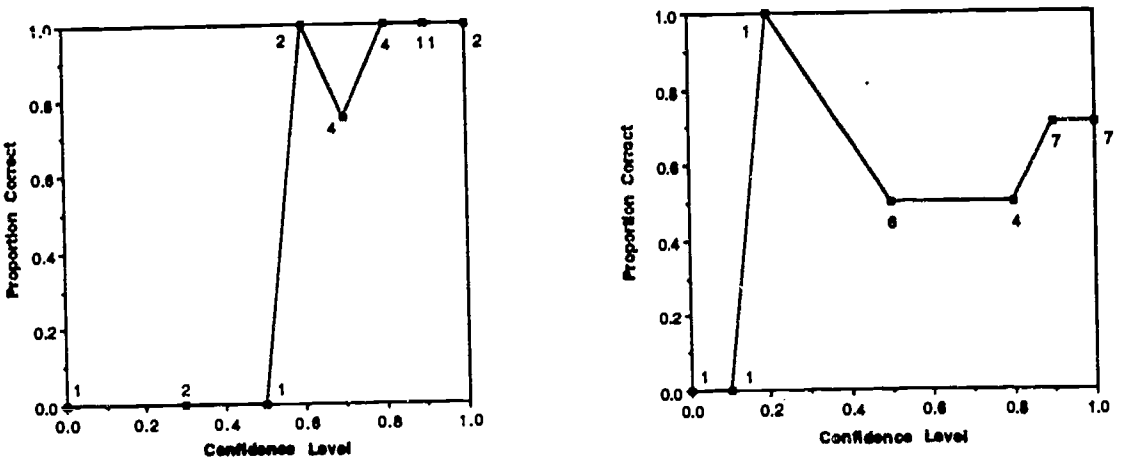


*Figure 2. Calibration curves for first-year student subjects #9 (left) and #2 (right).*

Subject #9 had excellent resolution. All four responses assigned confidence levels of 0%, 30% or 50% were incorrect diagnoses. Twenty-two of the 23 responses assigned subjective probabilities of 60-100% were correct diagnoses. In contrast, subject #2 had poor resolution. Knowing this subject's confidence level provides little information as to whether a particular diagnosis was correct or incorrect. The adjusted resolution of subject #9 in Fig. 2 was 0.82, but only 0.18 for subject #2.

Figure 3 illustrates the relationship between adjusted resolution and proportion correct for the two studies using second-order polynomial regression to fit a curve to the data.

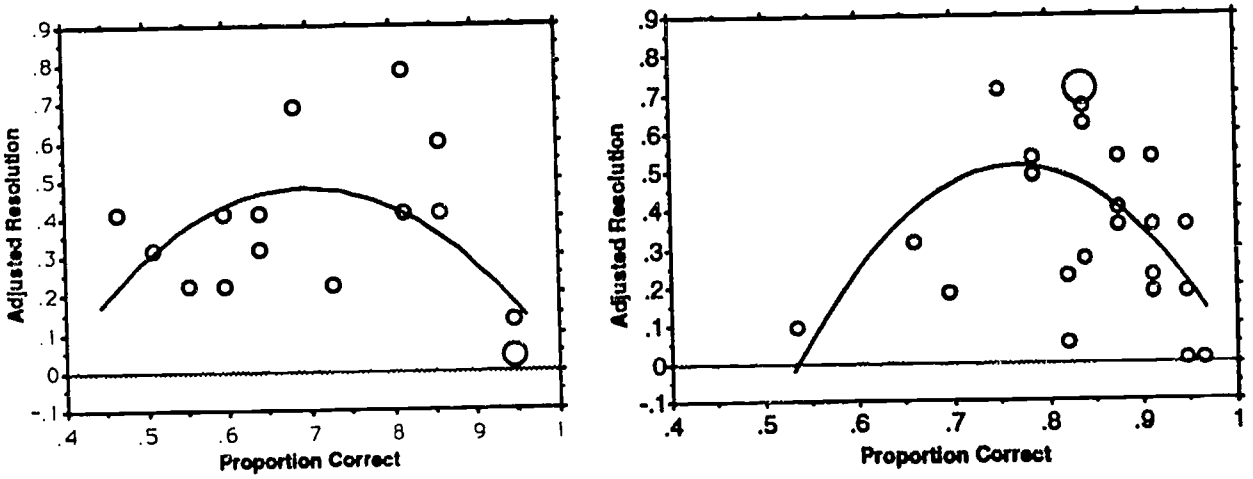


*Figure 3. Graphs of adj. resolution vs. prop. correct for first-year (left) & third-year (right) students*

For the 16 first-year students, the polynomial regression was not significant, $F(15) = 2.996$, $p = .087$ ($R^2 = .31$). For the 27 third-year students, the polynomial regression was significant, $F(24) = 5.058$, $p = .016$ ($R^2 = .32$).

## Discussion

In the limited domain of the classification of cardiac dysrhythmias, medical students are slightly underconfident rather than overconfident in their diagnostic ability. Their calibration curves look much like those of the "best subjects on easy items" in Lichtenstein and Fischhoff's (1977) experiments, which revealed substantial underconfidence at lower confidence levels and slight overconfidence at high levels of confidence. Further research will be required to distinguish between two possible causes for the medical students' underconfidence: their respect for the difficult domain of cardiology may have caused them to underestimate their ability in the relatively easy sub-domain of basic dysrhythmias, or they may simply not have acquired the sense of familiarity with the domain that leads some experts into overconfidence.

These two studies using adjusted resolution as a measure of metacognition replicate the findings of Lichtenstein and Fischhoff (1977) using calibration to assess metacognition, with the highest level of metacognition achieved by subjects at 70-85% accuracy. Subjects at lower accuracy levels might be expected to be less able to reflect on their own performance than more knowledgeable subjects. However, many subjects at accuracy levels above 85% appear to have a reduced awareness of the accuracy of individual diagnoses, which is a disturbing finding. These subjects' demonstrated expertise should enable them to achieve the highest adjusted resolution scores rather than some of the lowest.

Recording response times may help to discover a relationship between automaticity and lower adjusted resolution. Automaticity and pattern recognition may develop rapidly in the classification of distinctive visual stimuli. Further research is also required to discover whether expertise in a domain, no matter how recently acquired, produces a "confidence set" that reduces a novice's (or an experienced physician's) ability to reflect on his or her diagnostic performance in individual cases.

These preliminary studies have implications for medical education. Medical students appear to differ more widely in diagnostic metacognition than in accuracy; in these studies, metacognition as measured by adjusted resolution varied from 0.02 to 0.82, while accuracy was always above 44%, with most subjects scoring above 70%. Diagnostic metacogntion is critical to prompt graduates to use the knowledge retrieval skills (e.g., literature searching) currently being taught at many medical schools; physicians must realize that they need more information before they can benefit from access to knowledge bases, expert consultants or computer-based diagnostic assistance systems.

To the extent that diagnostic metacognition is considered important and is influenced by the initial acquisition of diagnostic categories during medical school, teaching and testing methods that reinforce high levels of metacognition should be adopted. Research is currently underway to determine if diagnostic practice that requires students to practice each step of the rules for classifying dysrhythmias produces better metacognition than diagnostic practice that allows students to classify dysrhythmias by whatever method they choose (e.g., rule shortcuts, pattern recognition). Assessing diagnostic ability using confidence-weighted response scoring should produce better diagnostic metacognition over a period of time. Building the basis for diagnostic metacognition in medical school may reduce the need for elaborate "debiasing" strategies that attempt to correct the systematic judgment errors of experienced decision-makers (Arkes, 1991).

## References

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110,* 486-498.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1-3.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 928-935.

Hausman, C. L., Weiss, J. C., Lawrence, J. L., & Zelenzik, C. (1990). Confidence weighted answer technique in a group of pediatric residents. *Medical Teacher, 12,* 163-168.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159-183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). New York: Cambridge University Press.

Mann, D. D. (1989). Effects of example analysis on diagnostic accuracy and confidence. Unpublished manuscript.

Mann, D. D., & Dane, P. B. (1991, April). The effect of practice cases and debiasing on rule-based diagnosis of similar cases. In J. Michael (Chair), *Applied studies in clinical reasoning.* Symposium conducted at the annual meeting of the American Educational Research Association, Chicago.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12,* 595-600.

Poses, R. M., Cebul, R. D., & Centor, R. M. (1988). Evaluating physicians' probabilistic judgments. *Medical Decision Making, 8,* 233-240.

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Decision Processes, 42,* 271-283.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30,* 132-156.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Zeleznik, C., Hojat, M., Goepp, C. E., Amadio, P., Kowlessar, O. D., & Borenstein, B. (1988). Students' certainty during course test-taking and performance on clerkships and board exams. *Journal of Medical Education, 63*, 881-891.

# Appendix

## Decompositions of the mean probability score ($\overline{PS}$)

(adapted from Yates, 1982)

$$\overline{PS}\,(f,d) = \left(\frac{1}{N}\right)\sum_{i=1}^{N}(f_i - d_i)^2$$

and

Mean Probability Score [$\overline{PS}$]

MURPHY DECOMPOSITION:

Outcome Index Variance: $[\bar{d}(1-\bar{d})] - [\left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j(\bar{d}_j - \bar{d})^2]$ + Murphy Resolution + Reliability-in-the-Small: $[\left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j(f_j - \bar{d}_j)^2]$

COVARIANCE DECOMPOSITION:

Outcome Index Variance: $[\bar{d}(1-\bar{d})]$ - Murphy Resolution: $[\left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j(\bar{d}_j - \bar{d})^2]$ + Forecast Variance: $[S_f^2]$ + Reliability-in-the-Large: $[(\bar{f}-\bar{d})^2]$ - 2x Forecast-Outcome Covariance: $[2F_{sd}]$ + Murphy Resolution: $[\left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j(\bar{d}_j - \bar{d})^2]$

where

$f_i$ = forecast of the outcome for the $i^{th}$ case (i.e., confidence level from 0-100% that the answer to that item was correct)

$d_i$ = outcome index of the $i^{th}$ case: $d_i = 1$ if the judgment is correct (i.e., the right answer); $d_i = 0$ if incorrect (the wrong answer)

J = the number of forecast (confidence) levels used

N = the number of cases (questions)