Which model or type of cochlear implant is better?

Mario A. Svirsky, Sara Gallant, and Elad Sagi

Laboratory for Translational Auditory Research

Department of Otolaryngology-HNS

New York University School of Medicine

- How can we tell if a given clinical intervention is better than another one?

- This question has crucial importance in the field of cochlear implants.

- Potential patients (or their parents) want to know whether one cochlear implant device results in better outcomes than the those provided by the competition.

- **The highest level of evidence: A prospective, randomized, double-blind study.**
- **Unfortunately, these studies are rare in the cochlear implant field, perhaps due to practical or cultural reasons.**
- **Here we will review two major studies that have been (wrongly) interpreted as suggesting that device "X" is better than device "Z".**

An interesting study that has been misinterpreted.

A Preliminary Comparison of Performance between Patients Fit with the CII Bionic Ear™ and Patients Fit with the Nucleus 3G System

Michael F. Dorman, Anthony Spahr
Arizona State University
Tempe, Arizona 85158

Karen I. Kirk
Indiana University School of Medicine, Indianapolis, Indiana 46202

Advanced Bionics, 2003

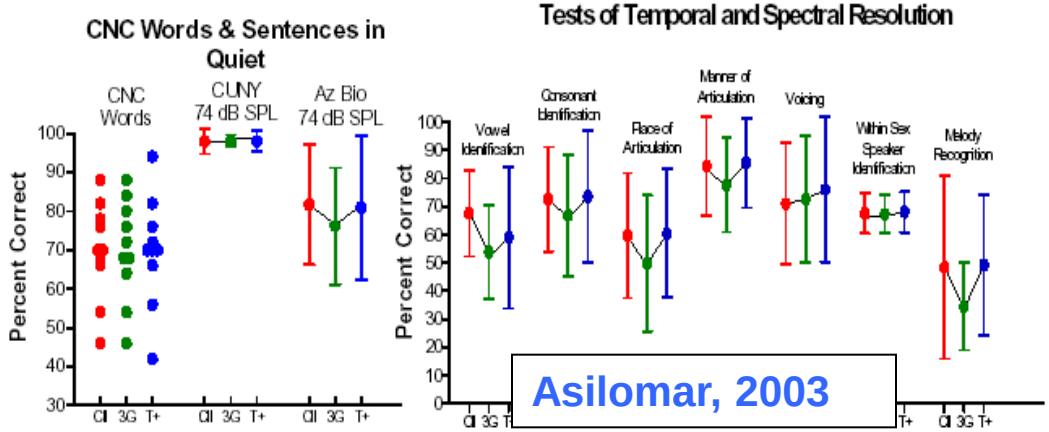
ORIGINAL ARTICLE

Performance of Subjects Fit With the Advanced Bionics CII and Nucleus 3G Cochlear Implant Devices

Anthony J. Spahr, MS; Michael F. Dorman, PhD

A COMPARISON OF PERFORMANCE AMONG PATIENTS FIT WITH THE CII HI-RESOLUTION, 3G AND TEMPO+ PROCESSORS

Anthony J. Spahr & Michael F. Dorman, Arizona State University



Asilomar, 2003

Main Outcome Measures: Vowel recognition, consonant recognition, sentences in quiet (74, 64, and 54 dB SPL [sound pressure level]) and in noise (+10 and +5

Arch Otolaryngol Head Neck Surg, 2004

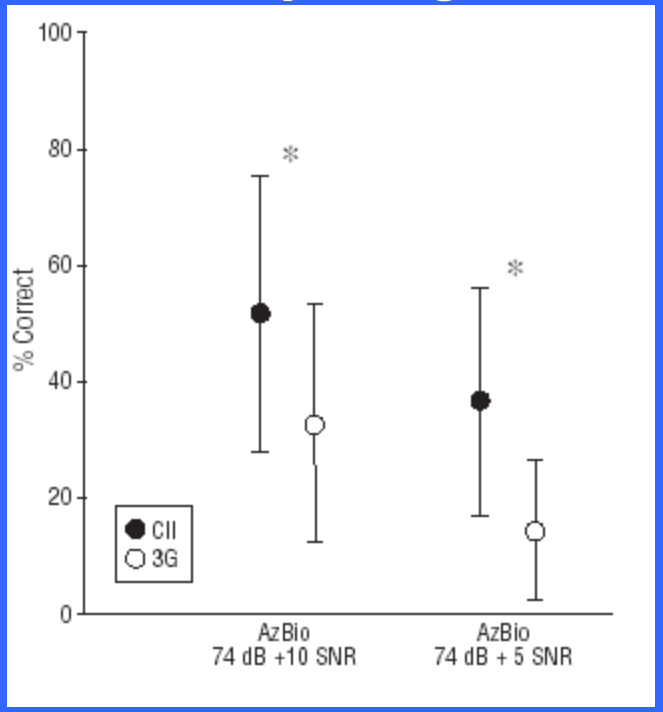
Conclusions: We have identified tasks on which there are between-group differences in performance for subjects matched on CNC word scores in quiet. We suspect that the differences in performance are due to differences in signal processing. Our next goal is to uncover the signal processing attributes of the speech processors that are responsible for the differences in performance.

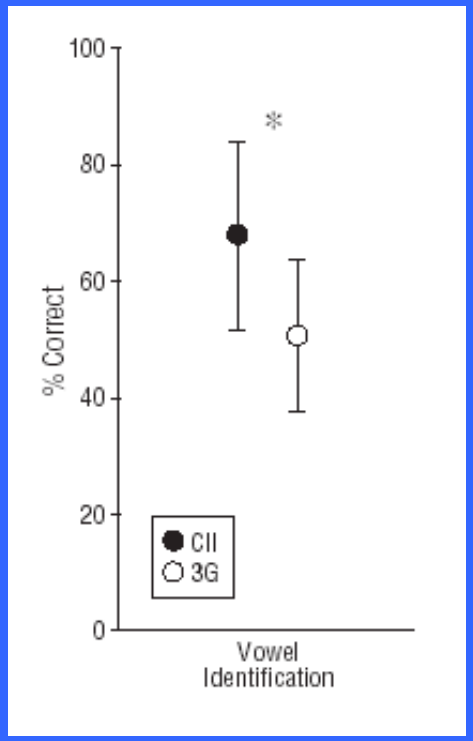
Arch Otolaryngol Head Neck Surg. 2004;130:624-628

- S&D: users of Clarion CII and Nucleus 3G devices were matched for CNC word scores and then compared in various other speech perception tests.

- Out of 16 tests, scores for AzBio sentences (+10 SNR, and +5 SNR), and vowel ID were higher for Clarion users.

- Important conclusion: when comparing CI devices we cannot trust any single test.





- Do these results support the hypothesis that the Clarion II is better than the Nucleus 3G? One manufacturer's web site seems to suggest this.

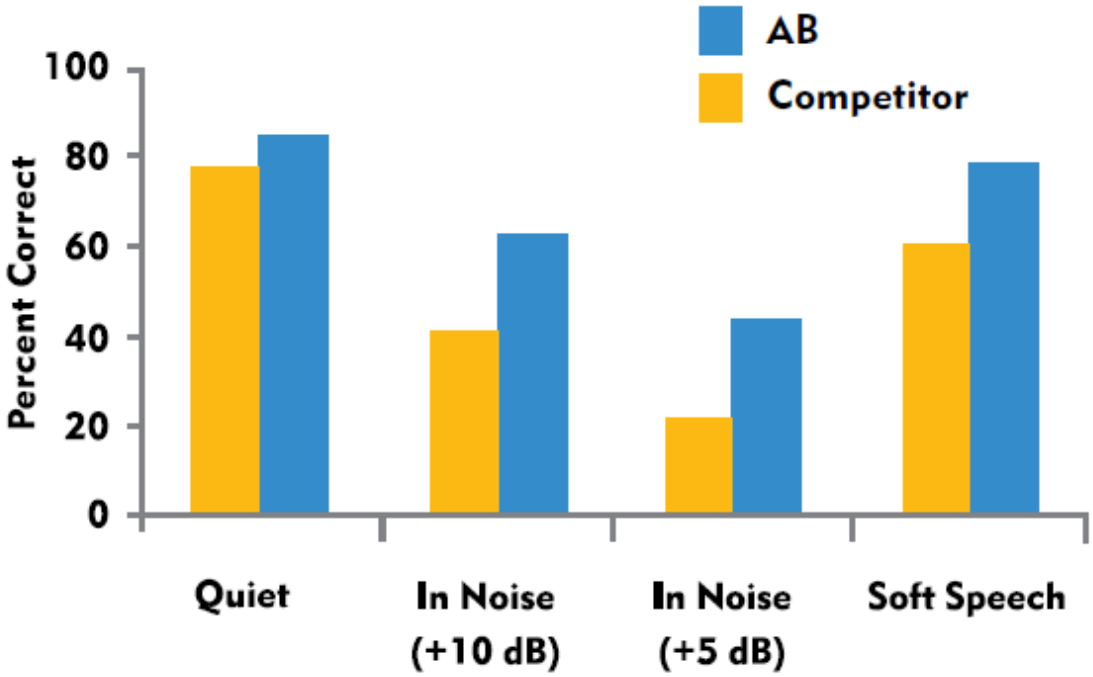
> Advanced Bionics quote:
>
> "…an independent study supported by an NIDCD grant and contributions from all CI manufacturers […] adults who use HiResolution sound and those who use the Nucleus 3 System. Results from the study's first reported findings indicate that there "are differences in performance between patients using the two implant systems." "Significant differences in performance ($p < 0.05$) were found between subjects in four test conditions and robustness in different listening environments."

Is this true?

What is "robustness"? Is it a good thing?

Real Science

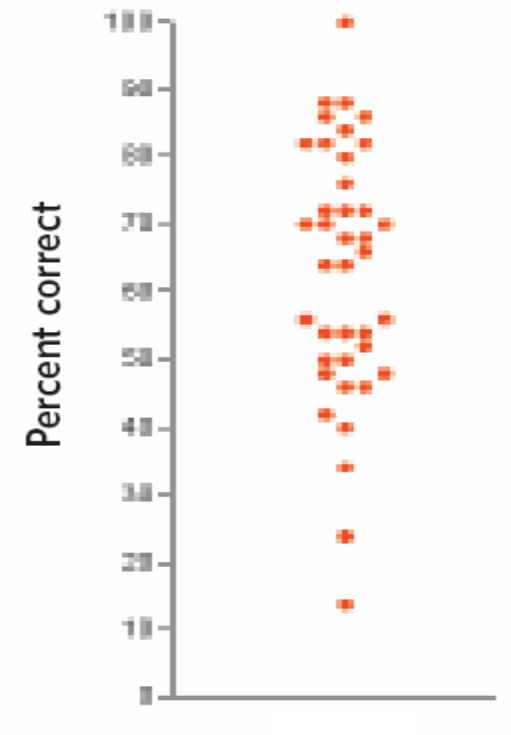
In an independent study by experts Spahr, Dorman, and Loiselle, AB outperformed the competition on tests of everyday listening.[1]



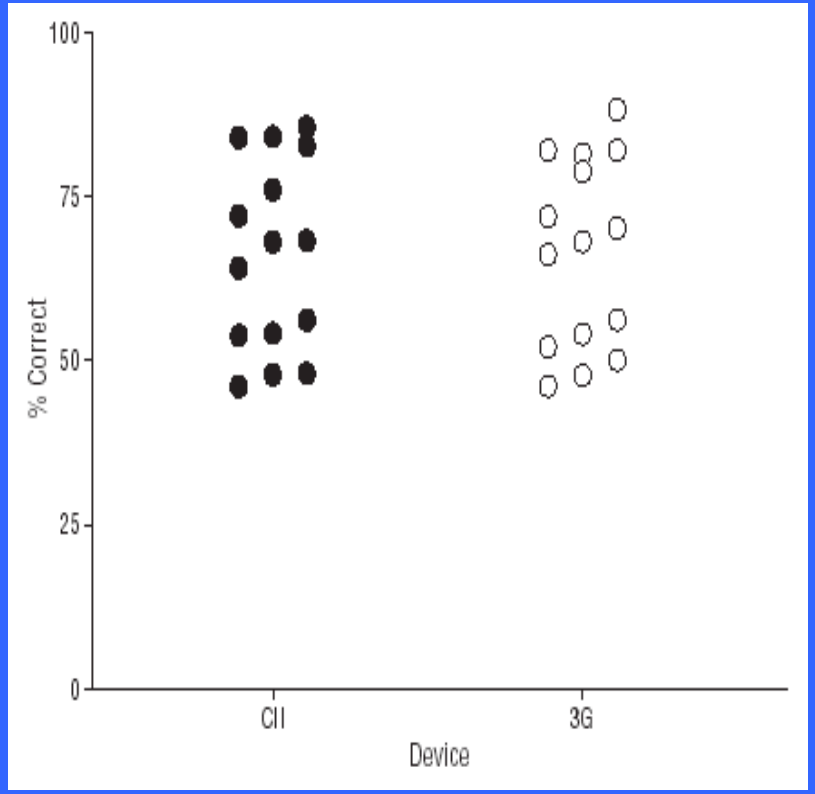
Mean sentence scores in quiet, noise, and at a soft level for 13 AB users and 13 competitor users. AB users hear better than competitor users in noise and when speech is soft.

Source:
"Hear Your Best Performance" Brochure, Advanced Bionics 2010

•S&D methods: Users of Clarion and Nucleus devices were tested for CNC words; patients with 40% or higher were matched and then compared in performance of other speech tests in various listening conditions

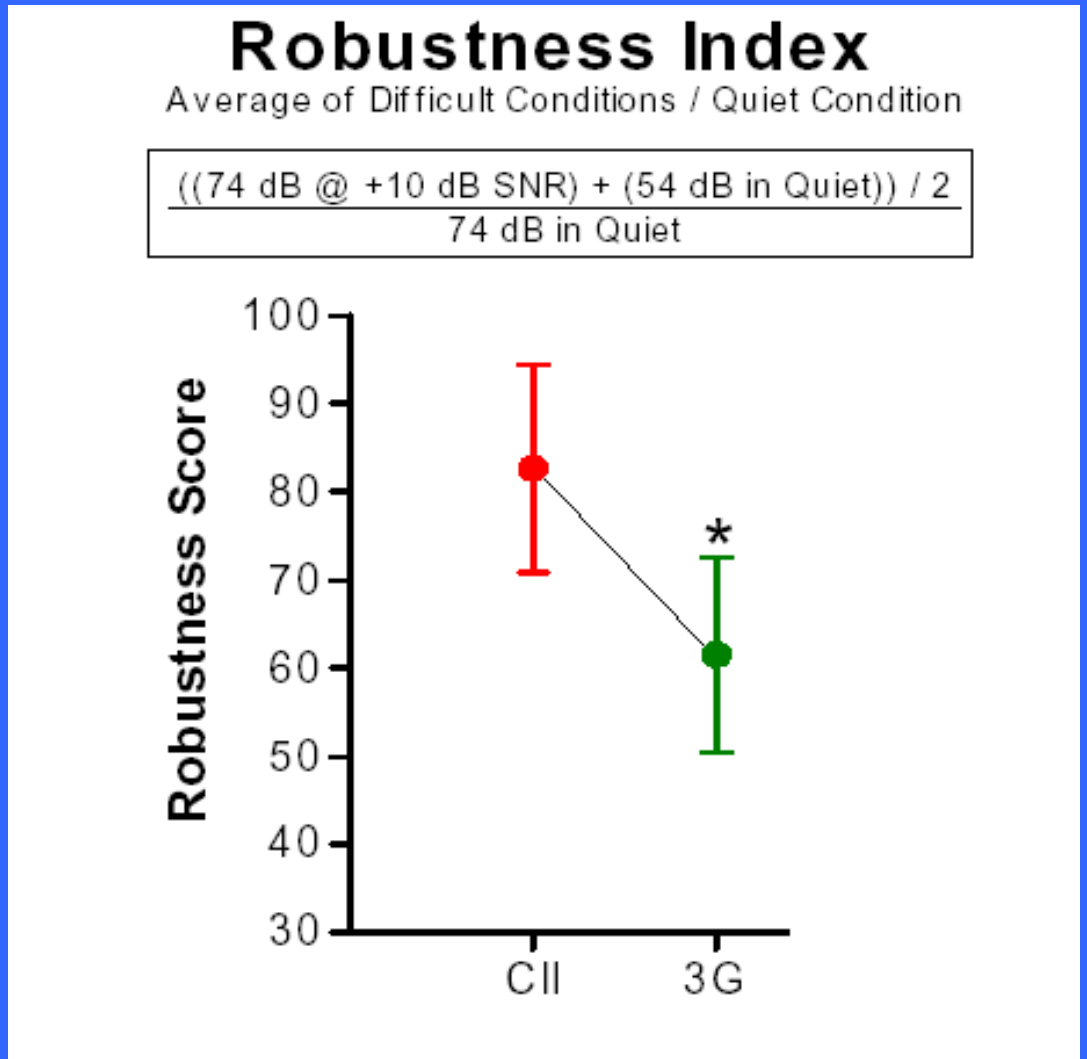


Overall CNC scores for both devices



CNC scores matched between 15 Clarion CII and 15 Nucleus 3G users

• Robustness: measures the difference in speech perception between “easy” and “difficult” conditions.



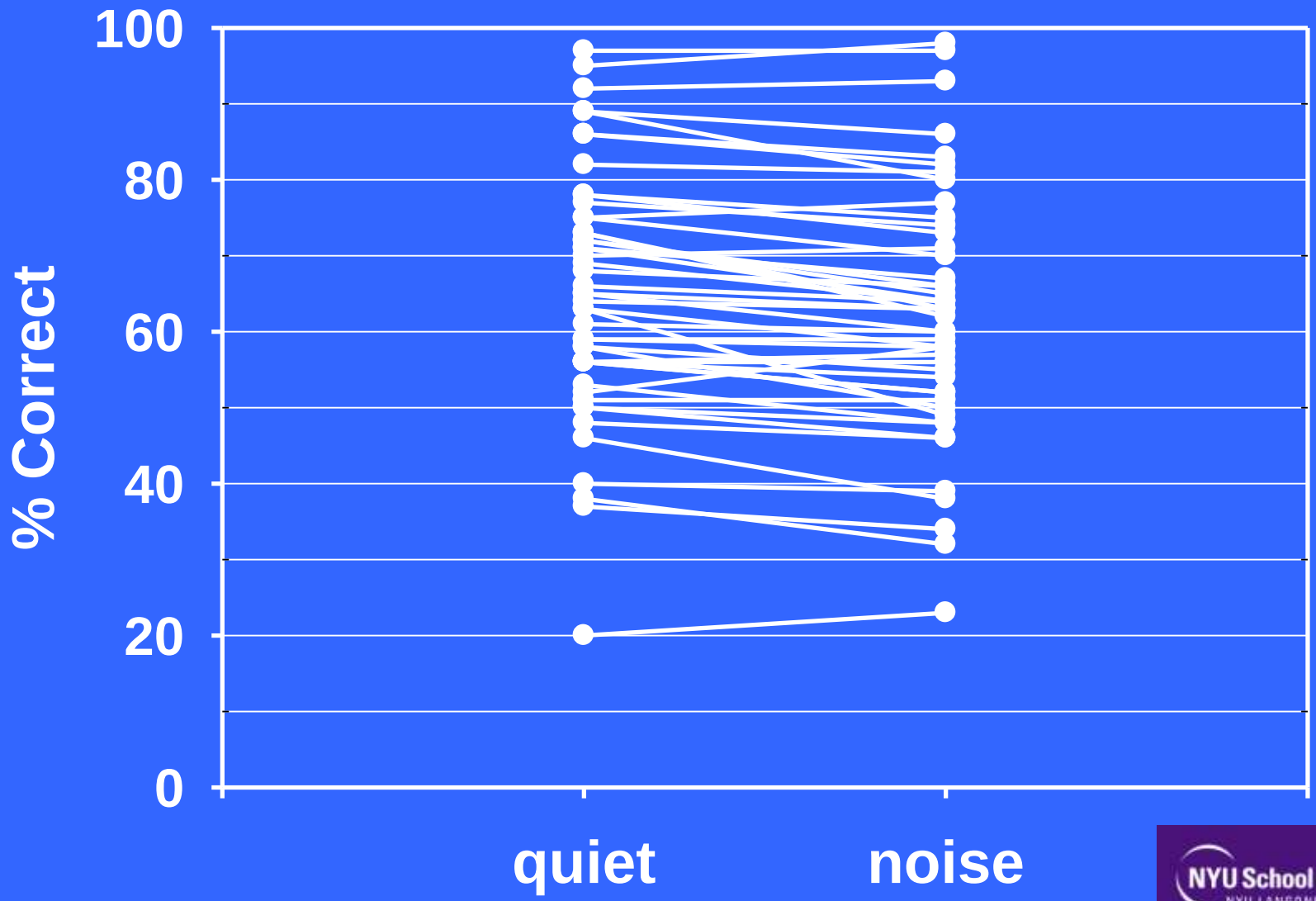
•Spahr and Dorman:

Clarion device has a higher robustness index than the Nucleus device

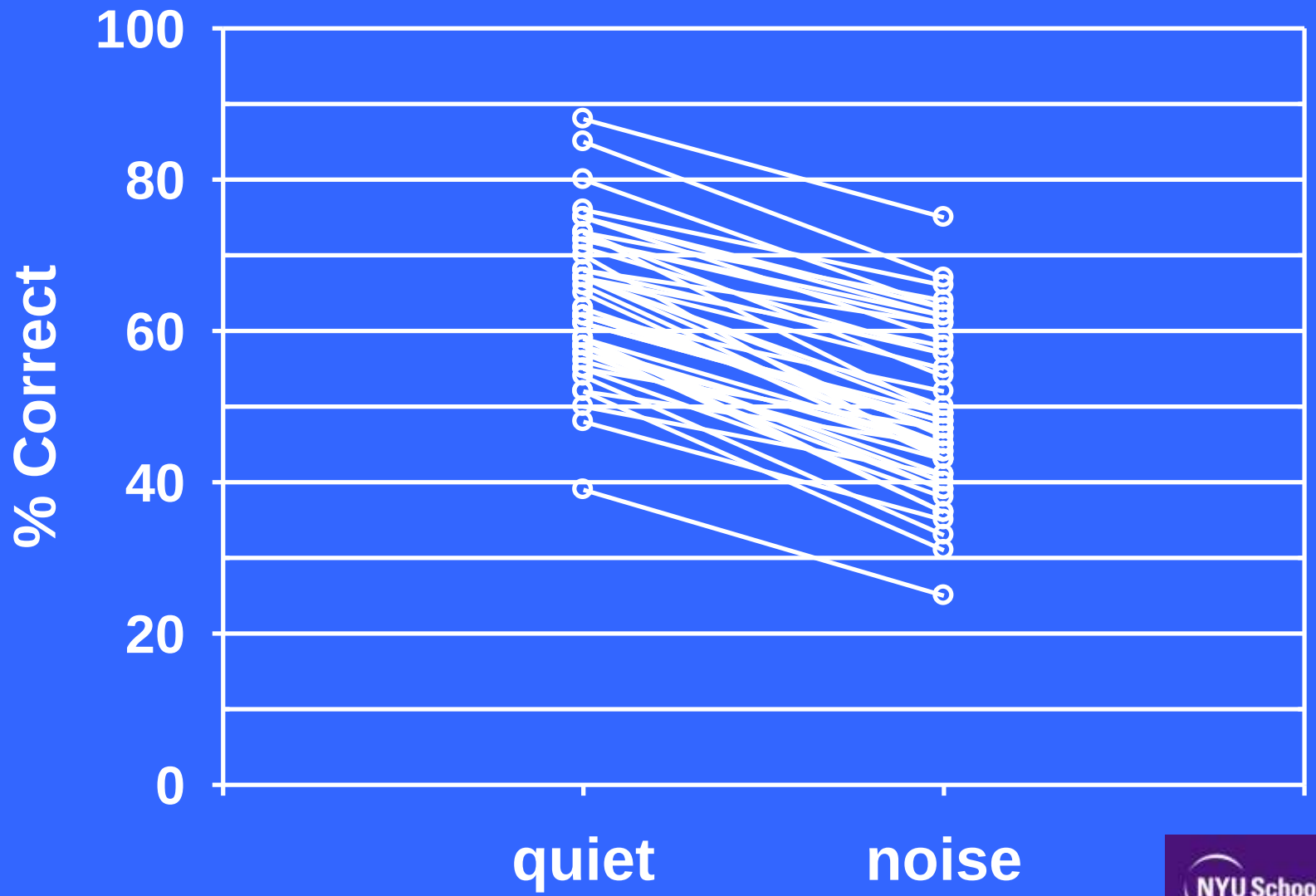
Methods of the present study

- A simulation study
- We generated pseudorandom scores for two devices (A and B), in quiet and in noise, in two "imaginary worlds".
- Scores for device A are about the same in noise and in quiet.
- Scores for device B are substantially better in quiet than in noise.
- In one imaginary world, device A is much better than device B. In the other imaginary world, B is better than A.

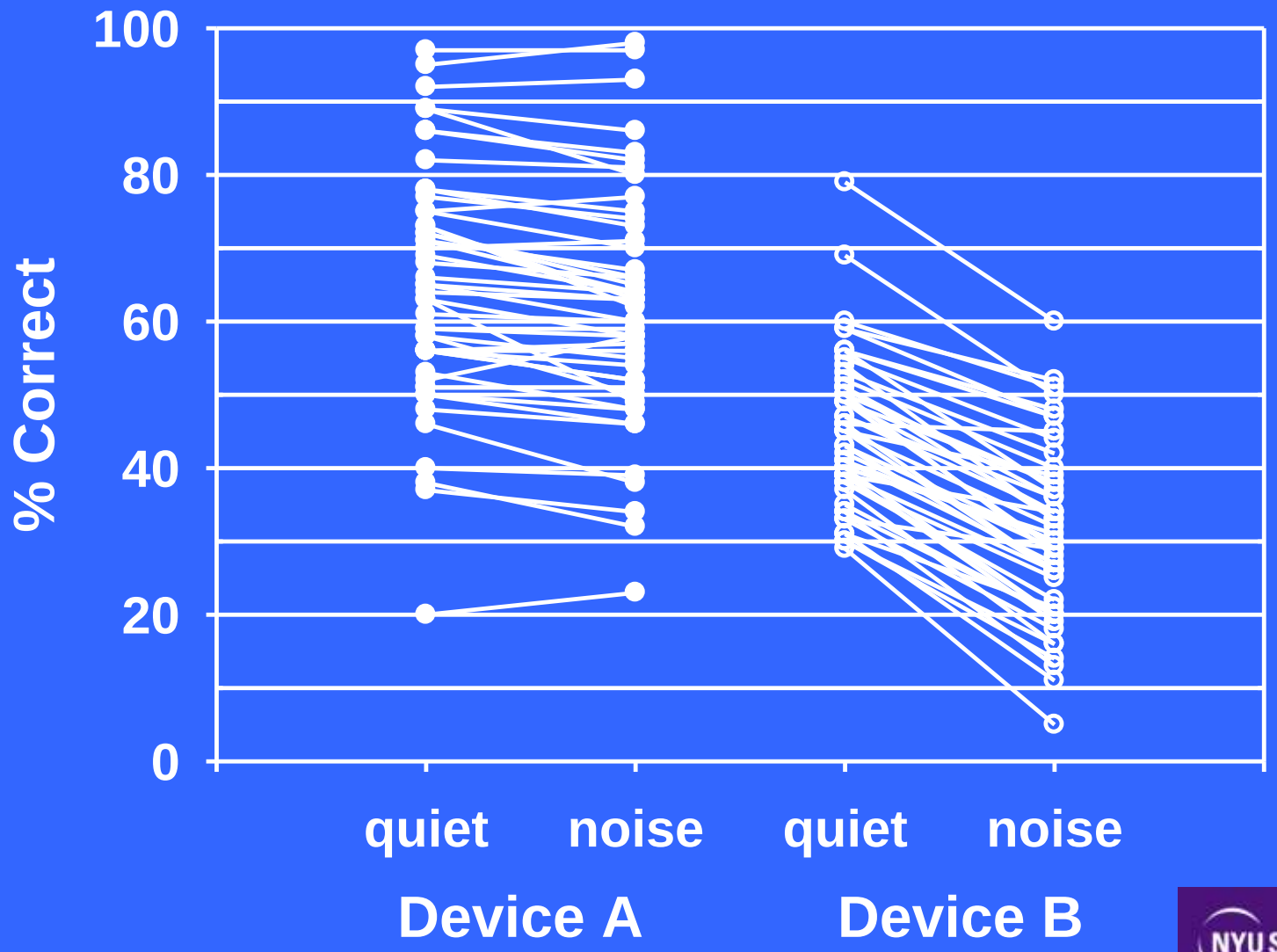
Device A: Similar performance in quiet and in noise



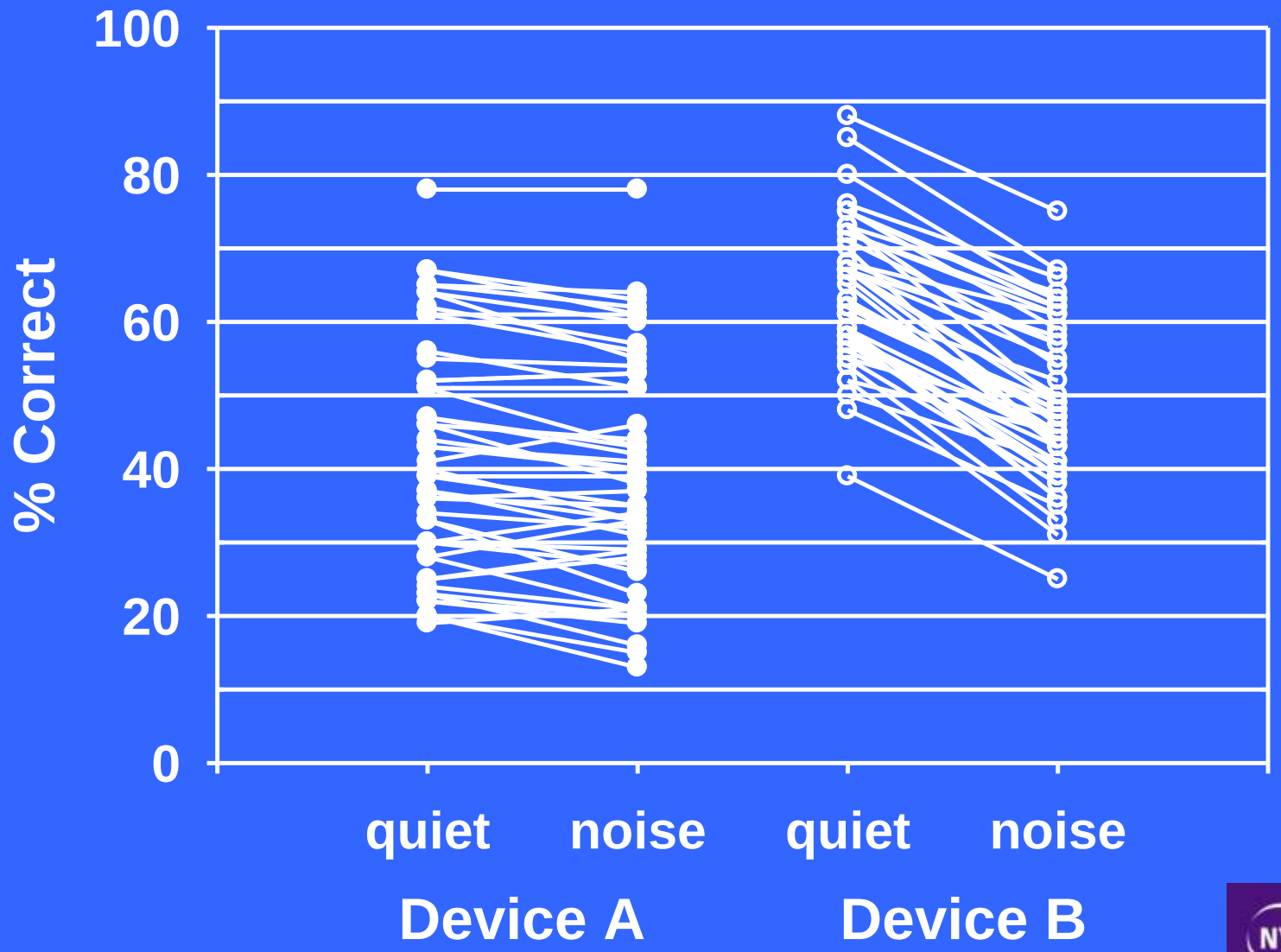
Device B: Much better performance in quiet than in noise



Imaginary World 1:
Device A is better than device B

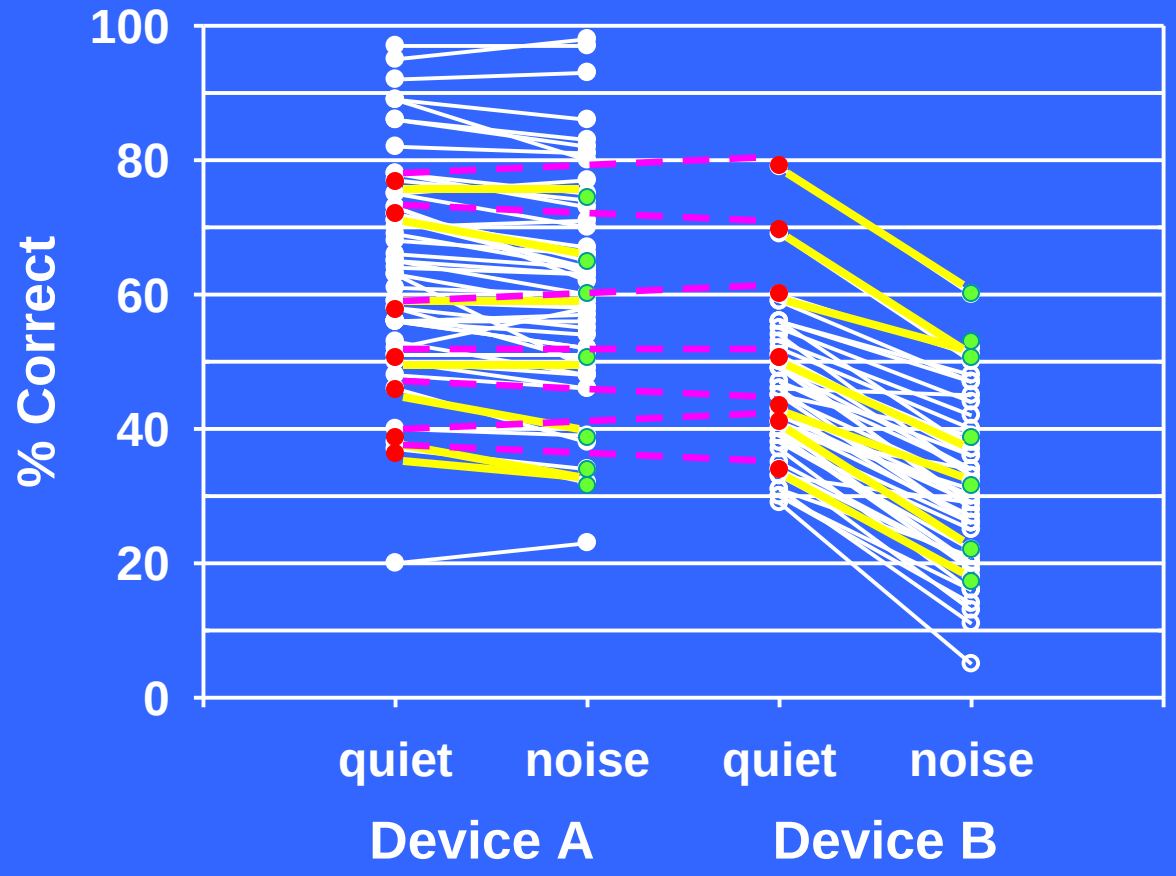


Imaginary World 2: Device B is better



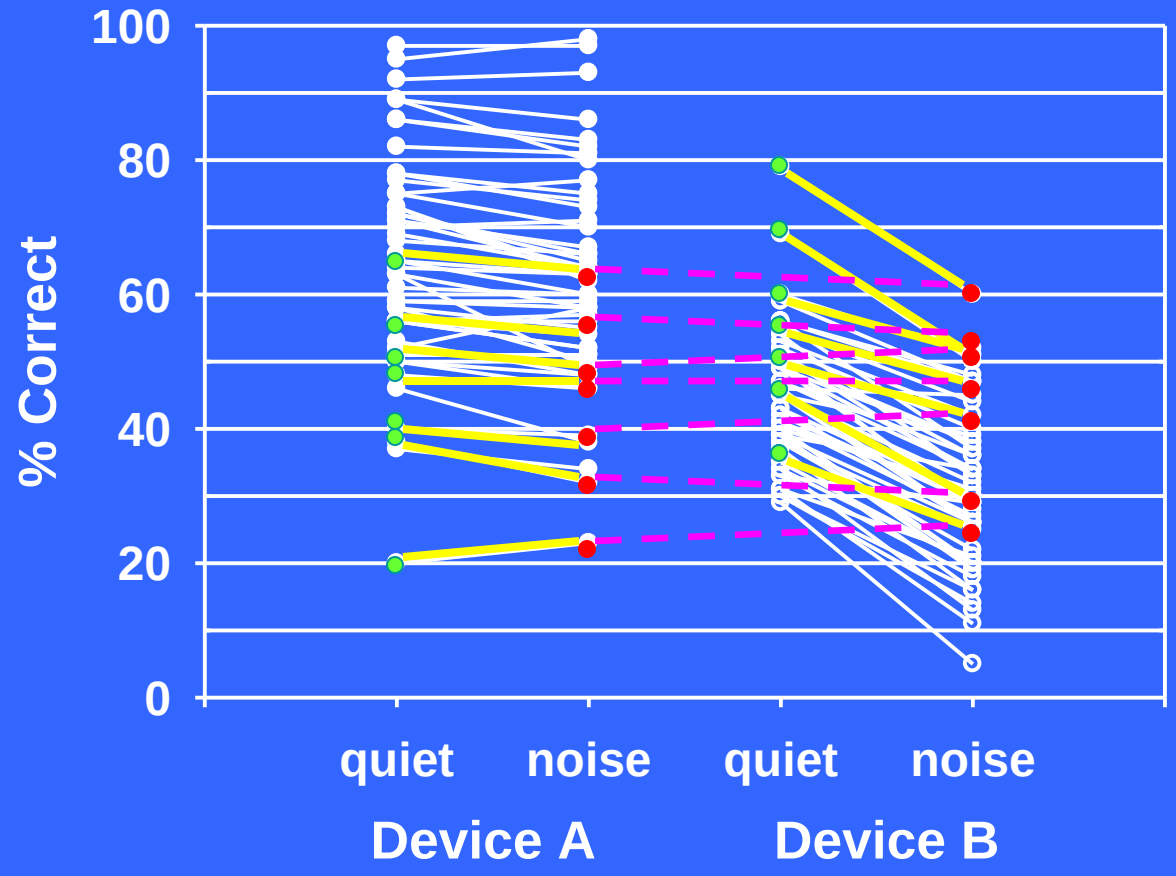
- Based on the simulated data from the two imaginary worlds, we conducted **two comparisons in each world**:
  - Matched pairs in quiet, compared their scores in noise (this is what Spahr and Dorman did).
  - Matched pairs in noise, and compared their scores in quiet.

Imaginary World 1 (device A is better)



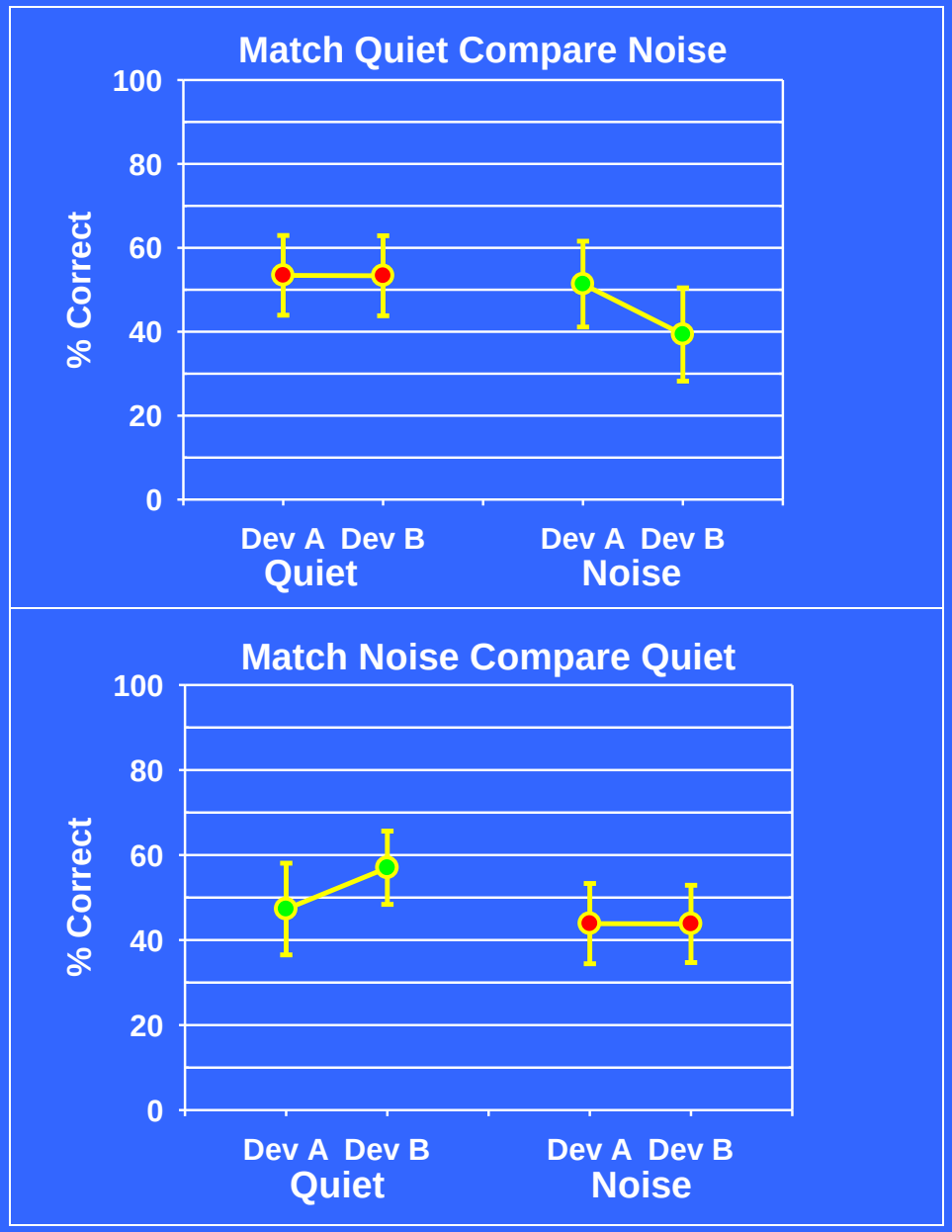
match in quiet, compare in noise

Still in imaginary World 1 (device A is better)

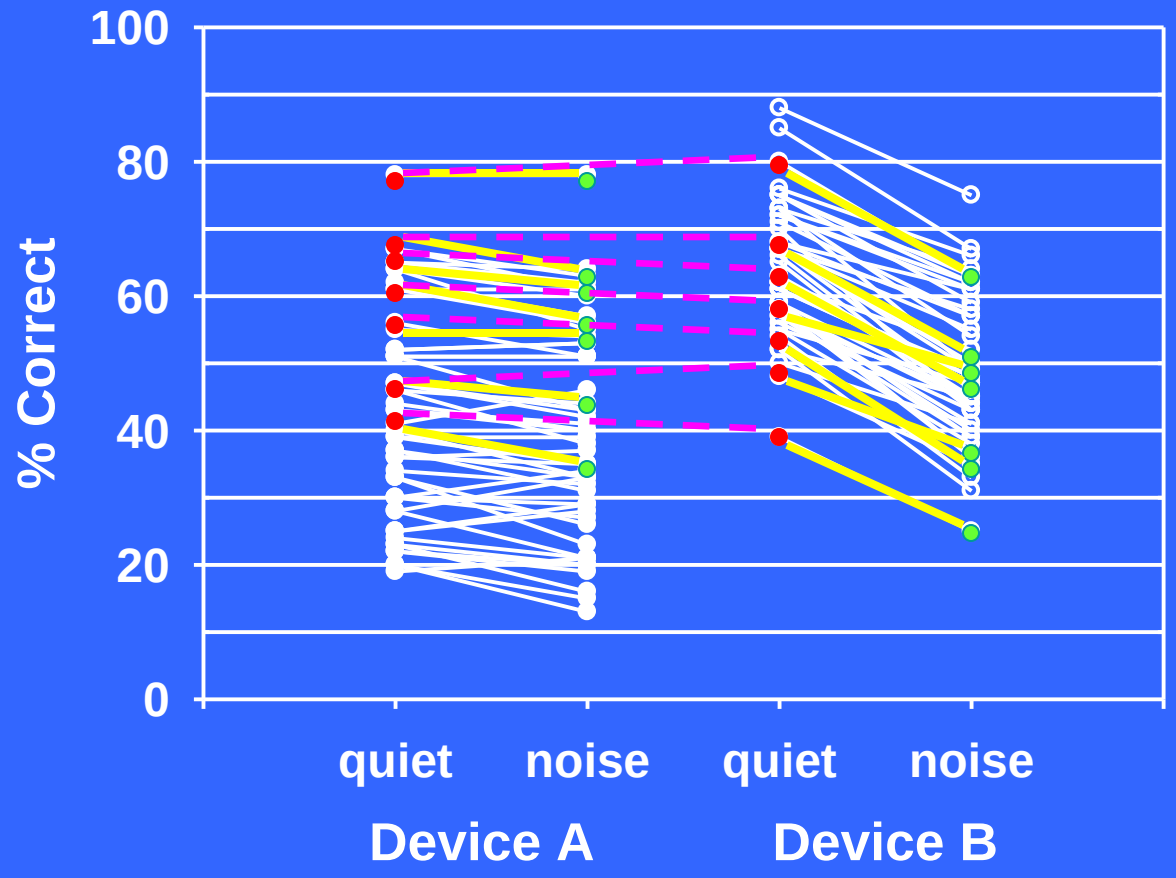


Now, match in noise and compare in quiet

Results of the two “World 1” comparisons

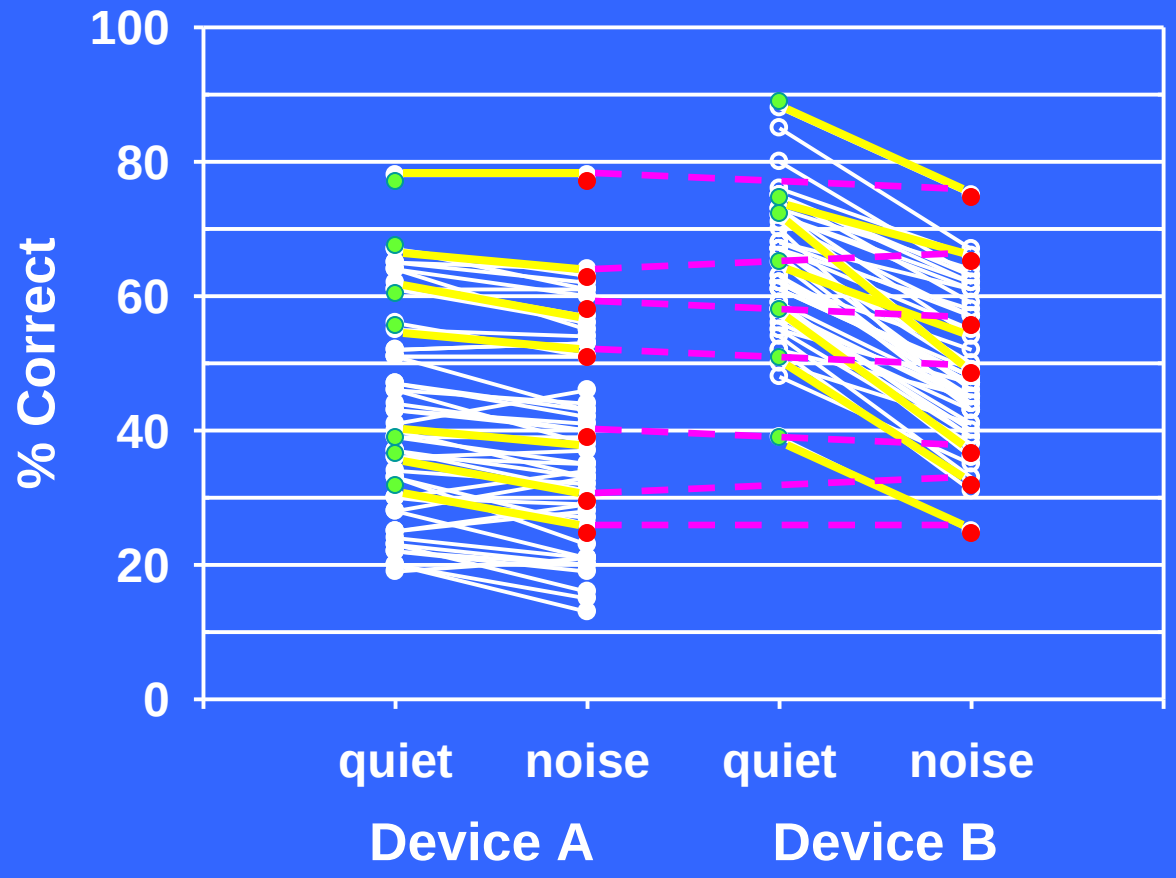


Imaginary World 2 (device B is better)



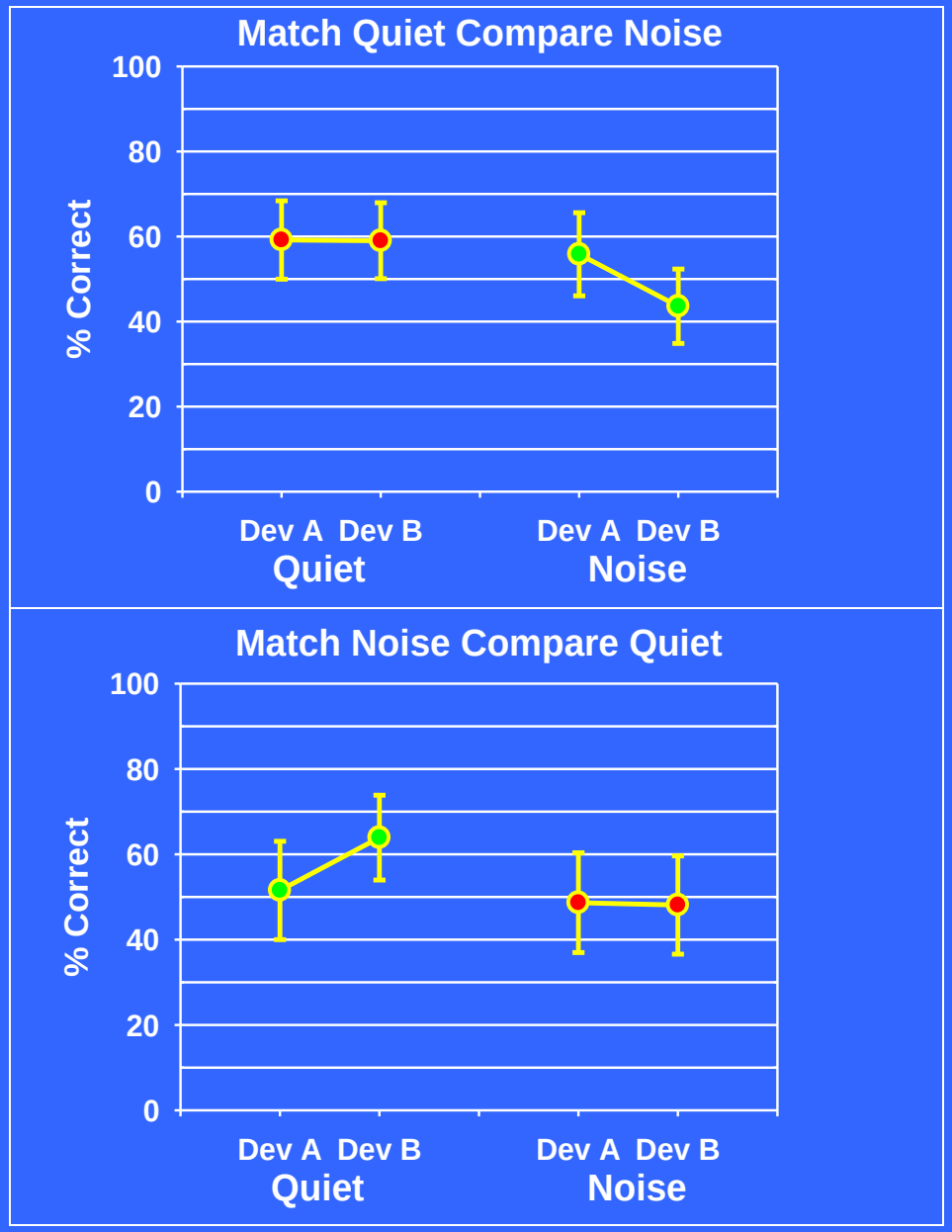
match in quiet, compare in noise

Still in Imaginary World 2 (device B is better)

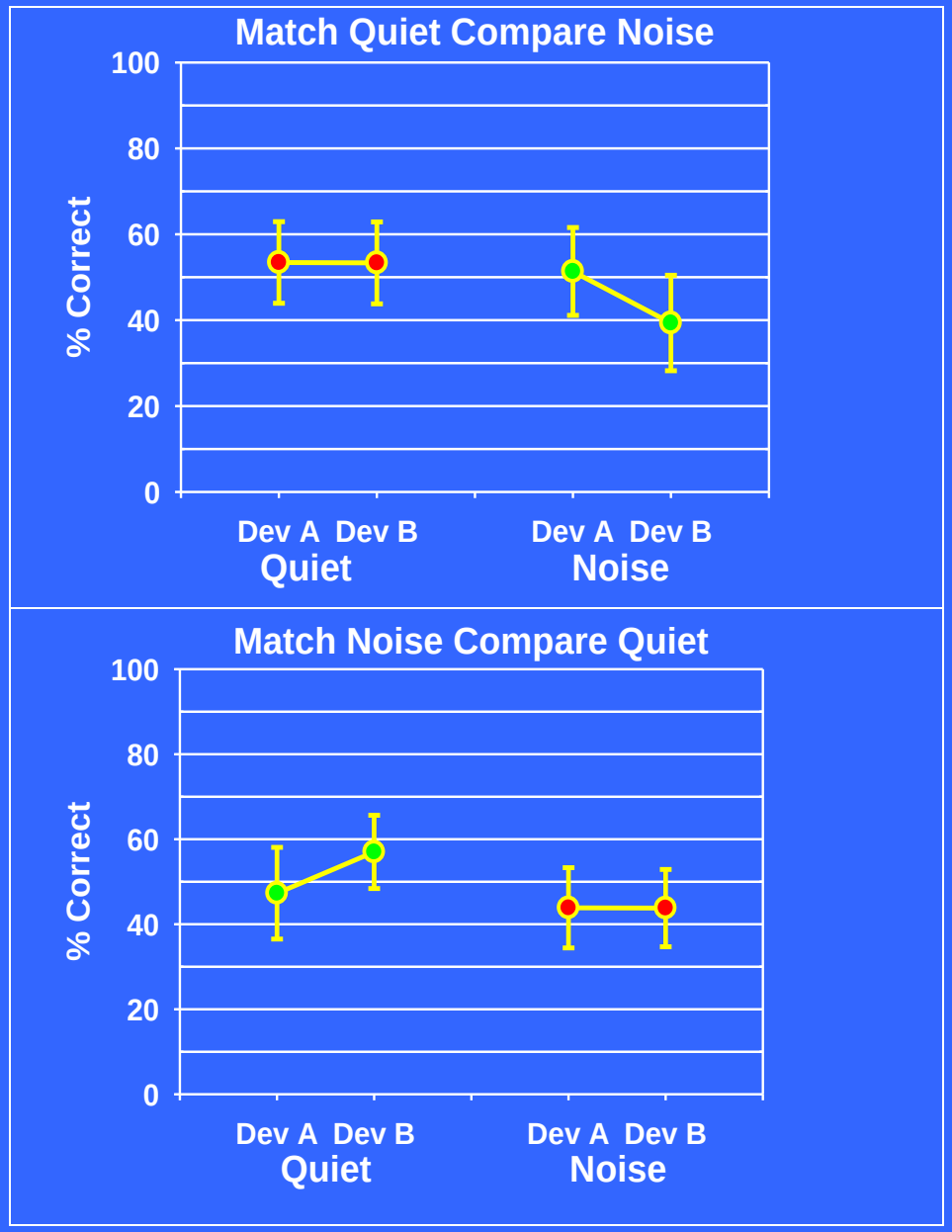


Now, match in noise and compare in quiet

Results of the two “World 2” comparisons



Results of the two “World 1” comparisons



- **Results are qualitatively identical, regardless of whether we use data from World 1 or World 2.**

- **When data are matched in quiet, device A has better results in noise.**

- **When data are matched in noise, device B has better results in quiet.**

- **Both these results happen in World 1 and in World 2.**

- **In conclusion, the matching procedure used in this study (and in the S&D study) cannot possibly tell us whether one device is better than another.**

What about the robustness index?

- The robustness index measures the difference between “easy” and “difficult” listening conditions.
- Robustness index=

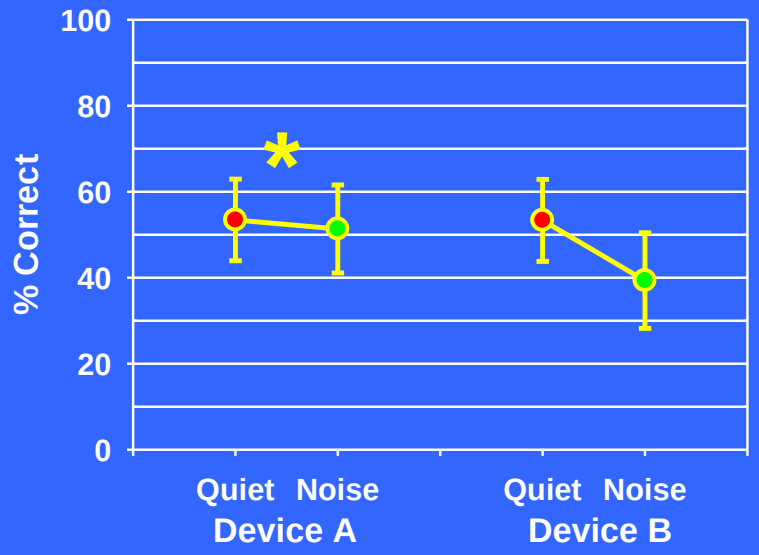
$$\frac{[(74\text{ dB in noise} + 54\text{ dB in quiet})/2]}{74\text{ dB in quiet}}$$

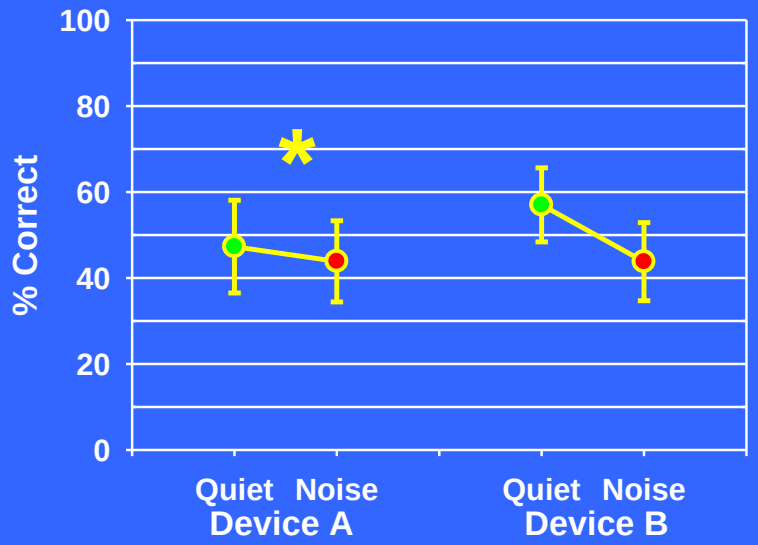
Robustness index in both worlds

Matching in quiet

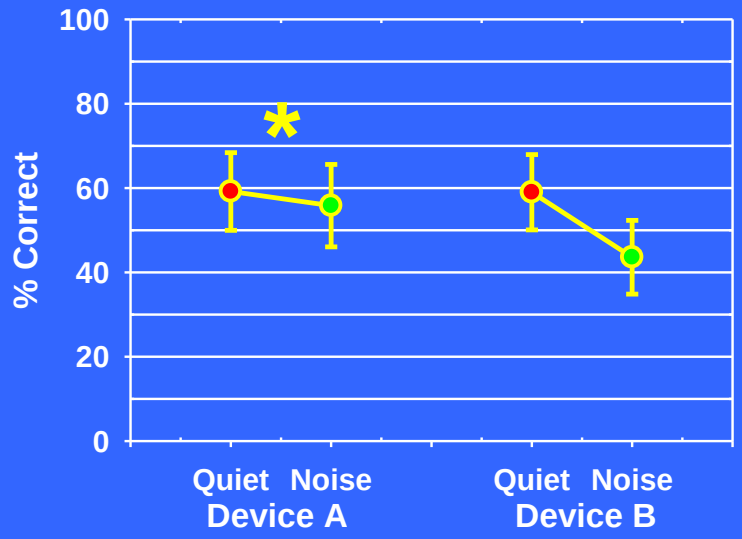
Matching in noise

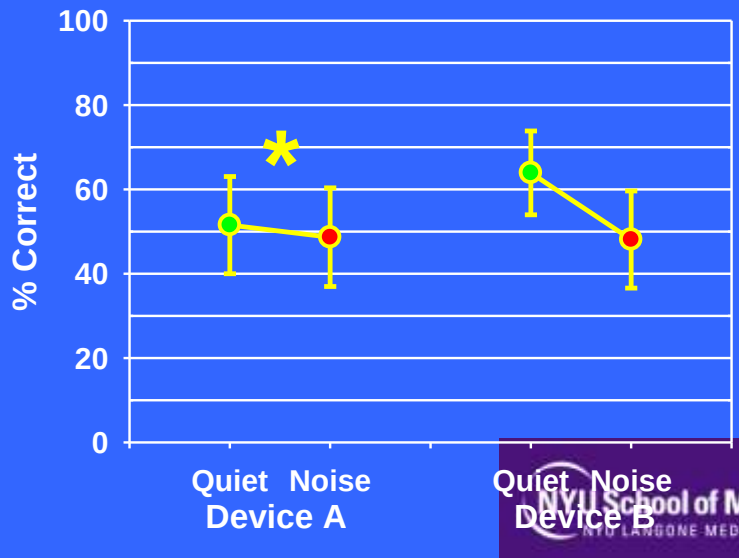
World 1



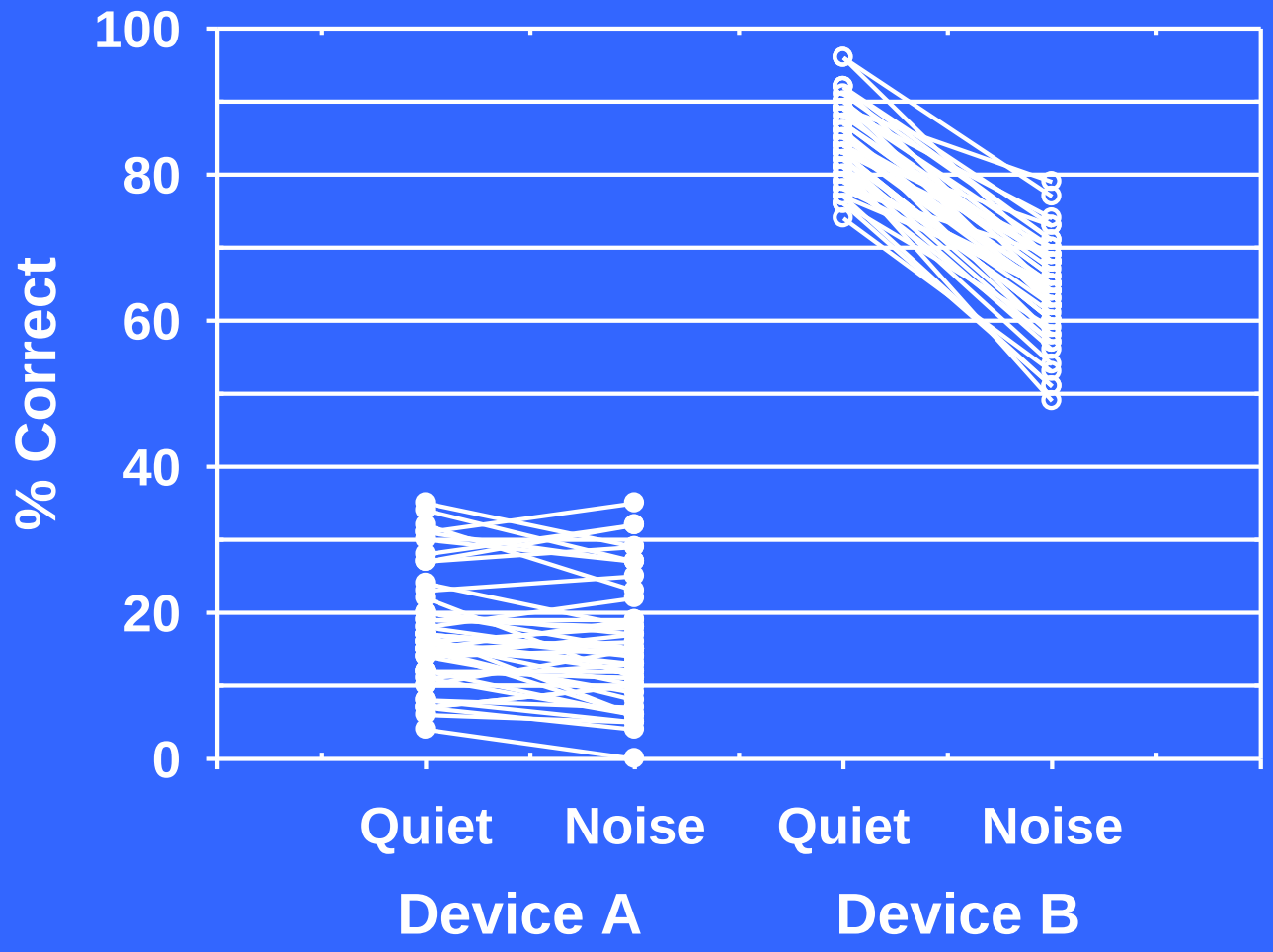


World 2



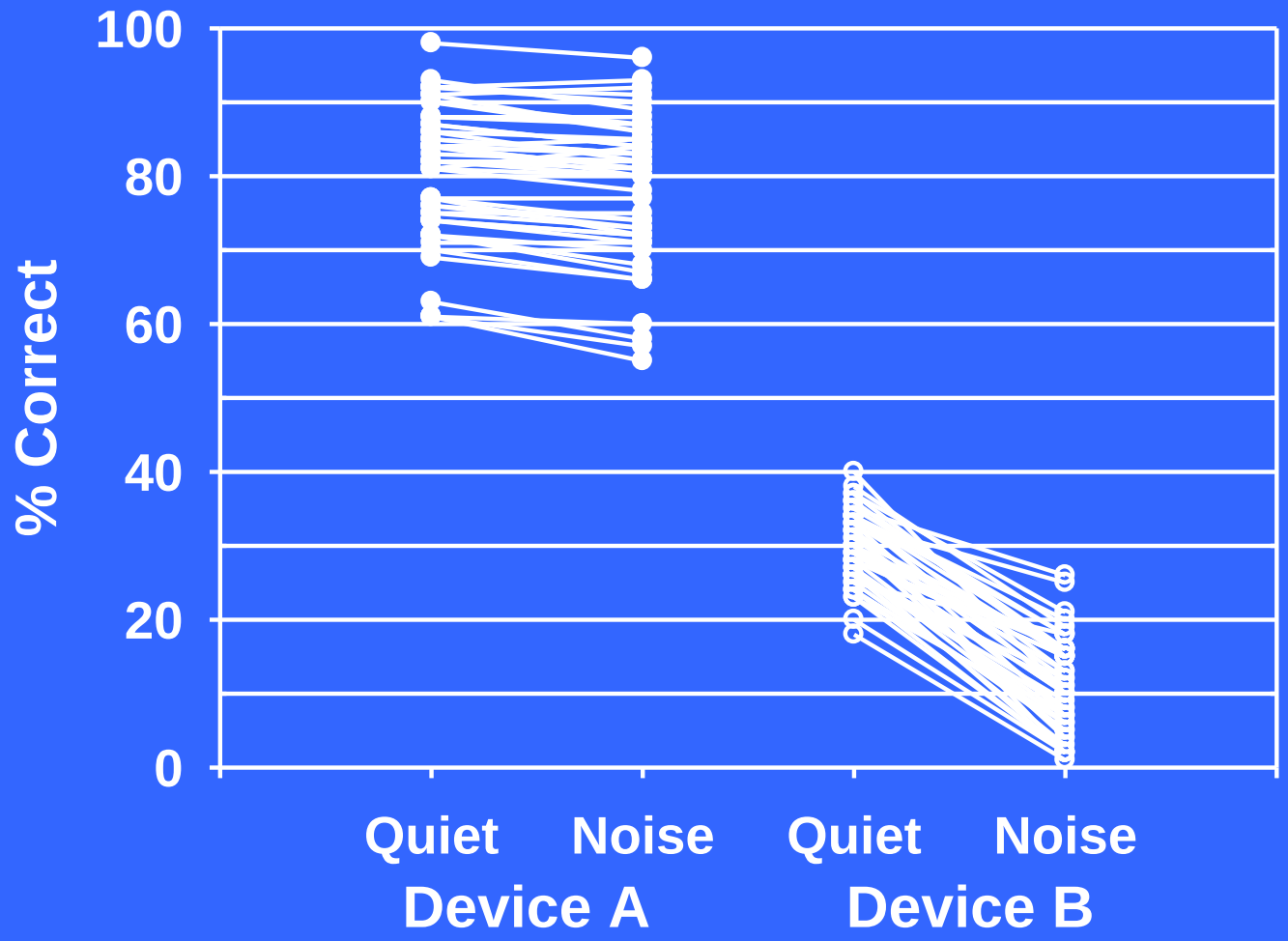


Extreme example 1: Device A has a higher robustness index…



…but it is much worse than device B

Extreme example 2: Device A still has a higher robustness index…



…but in this example it is much better than device B

Is a high robustness index good?

- Not necessarily.
- A device with a high robustness index may be much better or much worse than a device with a lower robustness index.

Conclusions

- The procedure employed in this study (matching subjects for word identification in quiet, then measuring speech perception in other tests) cannot be used to determine whether one cochlear implant is better than another. It is completely useless for that purpose.

- A device with a high robustness index is not necessarily better or worse than a device with a low robustness index.

Does this matter?

- Questionnaire
  - Administered to audiology students
  - 54 students filled out questionnaire a week apart, before and after reading 2004 Spahr and Dorman paper that used outcome-matching method to examine speech perception in noise and in quiet with those two cochlear implant devices
  - Asked to provide impressions about the performance of each tested CI device with respect to speech perception

Does this matter?

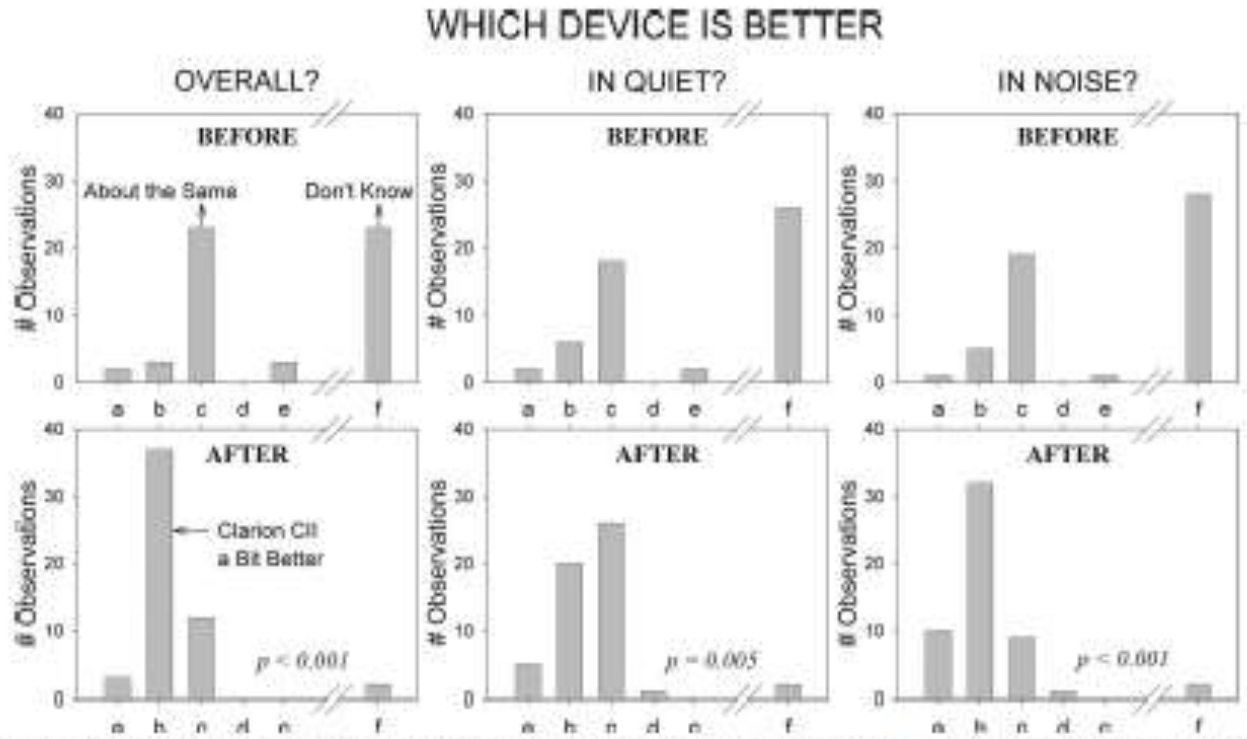


Fig. 5. Questionnaire results from 54 students before and after (upper and lower panels, respectively) reading Spahr and Dorman (2004). Respondents were asked for there impressions of the Clarion versus Nucleus devices in terms of overall speech perception, speech perception in quiet, and speech perception in noise (left, middle, and right columns, respectively). The *p* values represent a paired comparison from before to after.

Does this matter?

- Questionnaire
  - Number of “don’t know” responses dropped significantly after reading the paper
  - Significant number of respondents changed their impression about the two devices

Does this matter?

– Conclusions

- Informed readers can easily misinterpret the results of the outcome-matching method by inferring that significant differences between matched groups generalize to the population as a whole

Is it possible to know for certain whether a given cochlear implant is better?

- Yes, by conducting a prospective, randomized, double blind study.

- This is difficult (particularly the "double blind" part) but not impossible. Precedent: the VA study (Waltzman and Cohen).

- Another possibility: using convenience samples of subjects, and trying to control the influence of covariates by matching, or by statistical procedures.

- This type of study represents a relatively low level of evidence.

- Lazard et al., 2012
  - Objective:
    - Test influence on CI speech performance in quiet and in noise for postlinguistically deaf adults
      - Duration of moderate hearing loss
      - Surgical approach
      - Angle of insertion
      - Percentage active electrodes
      - Brand of device
      - Duration of profound hearing loss
      - Age
      - Etiology
      - Duration of CI experience

- Lazard et al., 2012
  - Objective:
    - Design model of predicted auditory performance with a CI as a function of the significant factors

- Lazard et al., 2012
  - Study
    - Data from 2251 patients implanted since 2003 in 15 international centres
    - Speech scores in quiet and in noise converted into percentile ranks

- Postoperative speech scores in quiet and in noise were transformed into percentile ranks for each patient within each centre
- For each clinic, distribution varied uniformly from 0-100
  - Best performers from each center had percentile rank close to 100
  - Poorest performers had percentile rank close to 0

- Ranked data of the centres were combined for global analysis
- “Using ranking removes differences in clinical practice without removing the relative differences between patients within each clinic”

- Combining such data across centers can be complicated…

Example

- Consider this example of a similar experiment
  - 2 different centers
  - SRS scores calculated for 10 CI users
    - Mixed group of brand X and brand Y CIs
  - These scores are then transformed to a rank from 1-10

Example - Data

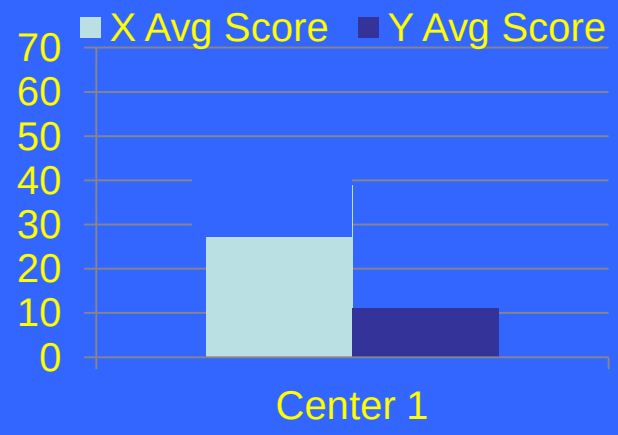
Brand	Score (%)	Percentile Rank
x	40	100
x	38	90
x	30	80
x	28	70
x	28	60
x	25	50
x	20	40
x	13	30
y	12	20
y	10	10

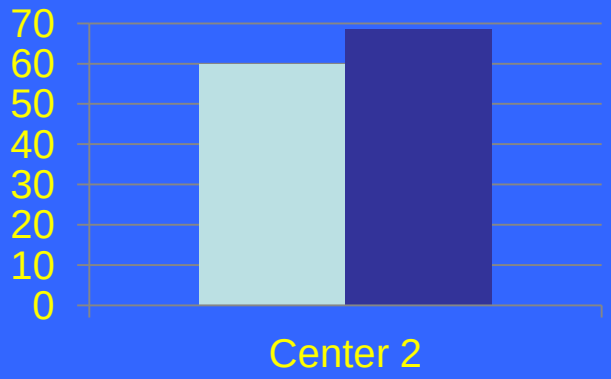
Center 1

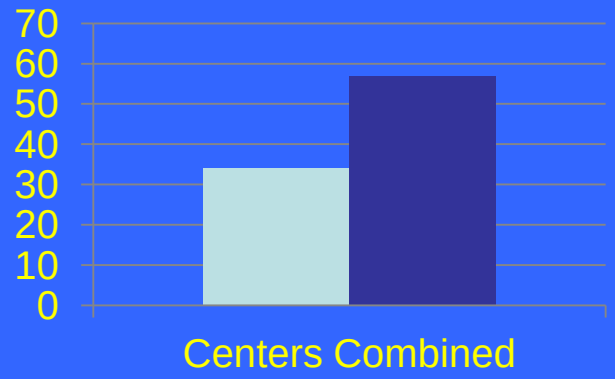
Brand	Score (%)	Percentile Rank
y	78	100
y	76	90
y	75	80
y	75	70
y	70	60
x	60	50
x	60	40
y	59	30
y	58	20
y	57	10

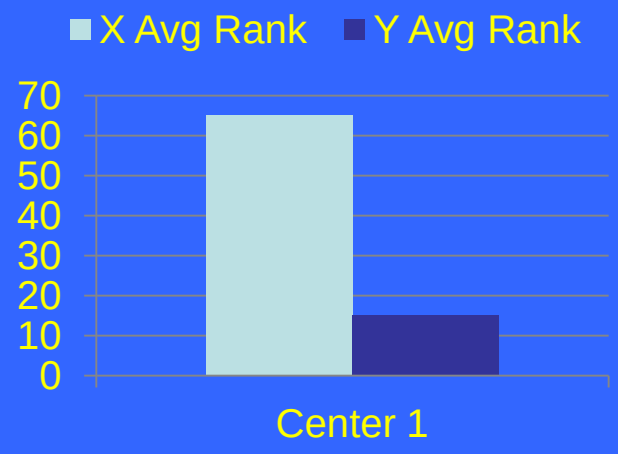
Center 2

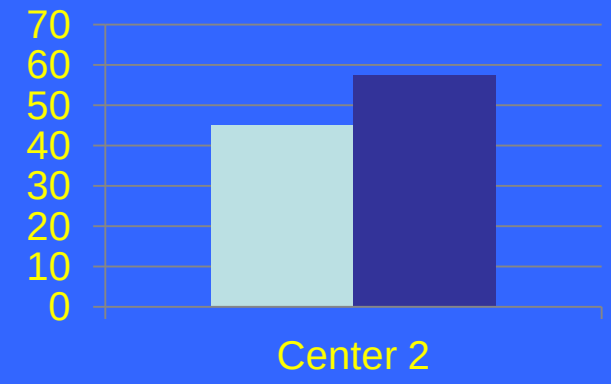
Example

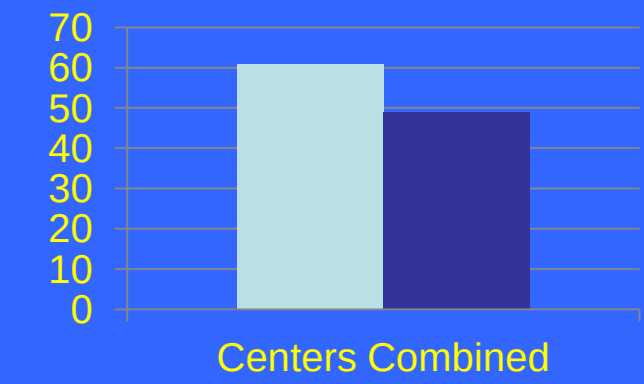






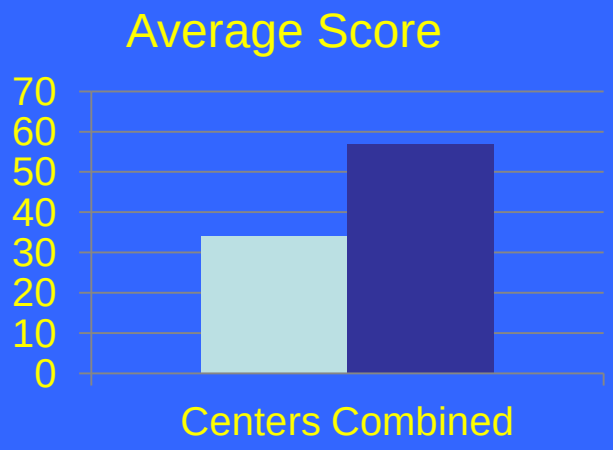


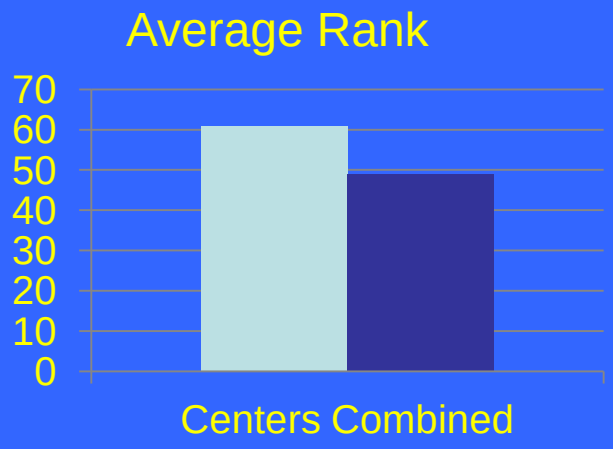




NYU School of Medicine
NYU LANGONE MEDICAL CENTER

Example





- When these data are lumped together across the two centers, the **better**-scoring implant (Y) has a **lower** rank

- Lazard et al., 2012
  - Study
    - Statistical analyses conducted on dataset
      - Each factor to be tested added into 4-factor analysis of variance using General Linear Model created by Blamey, which included
        - » Duration of severe/profound hearing loss
        - » Age of onset of severe/profound hearing loss
        - » Duration of CI experience
        - » Etiology
      - Yielded fifteen 5-factor ANOVAs
      - Factors with $p < 0.001$

- Lazard et al., 2012
  - Results
    - Significant factors:
      - PTA threshold of better ear
      - Percentage of active electrodes
      - Use of hearing aids during period of profound hearing loss
      - Duration of moderate hearing loss
      - And…

- Lazard et al., 2012
  - Results
    - Brand of device was significant
    - p = 0.000
    - Horizontal line represents average performance (5oth percentile for scores in quiet)

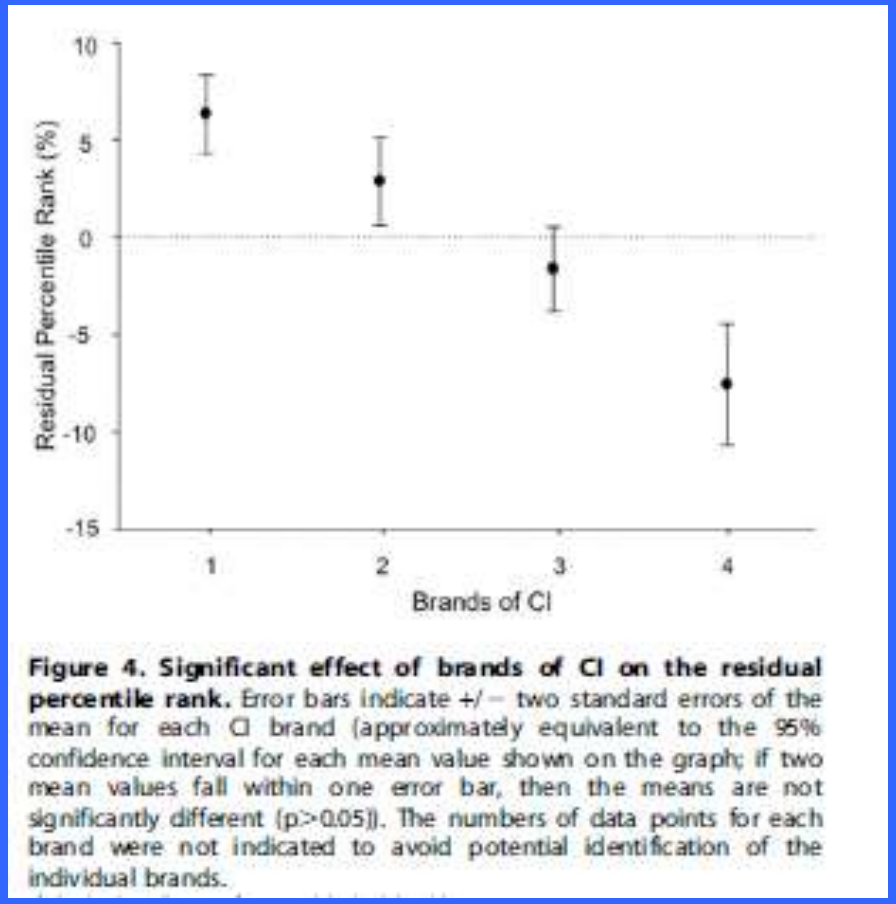


Figure 4. Significant effect of brands of CI on the residual percentile rank. Error bars indicate +/− two standard errors of the mean for each CI brand (approximately equivalent to the 95% confidence interval for each mean value shown on the graph; if two mean values fall within one error bar, then the means are not significantly different ($p > 0.05$)). The numbers of data points for each brand were not indicated to avoid potential identification of the individual brands.

- Lazard et al., 2012
  - Results
    - Difference between mean percentile rankings of highest and 2 lowest brands was significant
    - Mean scores of highest and lowest brands differed by only 14%

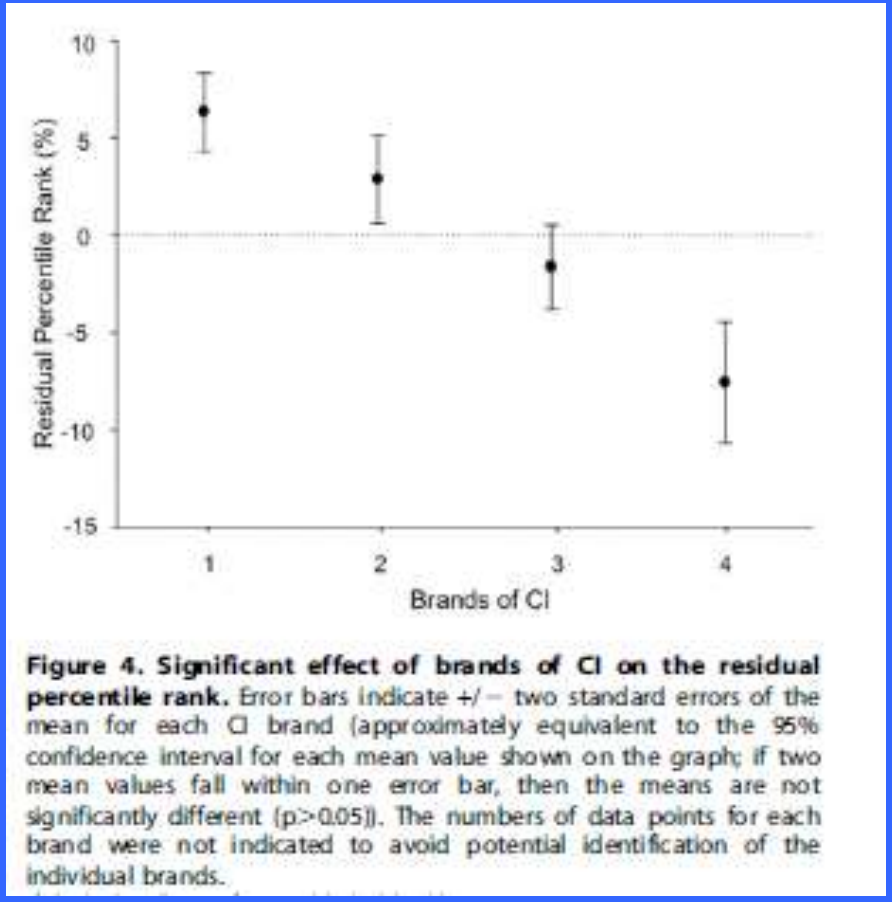


Figure 4. Significant effect of brands of CI on the residual percentile rank. Error bars indicate +/− two standard errors of the mean for each CI brand (approximately equivalent to the 95% confidence interval for each mean value shown on the graph; if two mean values fall within one error bar, then the means are not significantly different ($p > 0.05$)). The numbers of data points for each brand were not indicated to avoid potential identification of the individual brands.

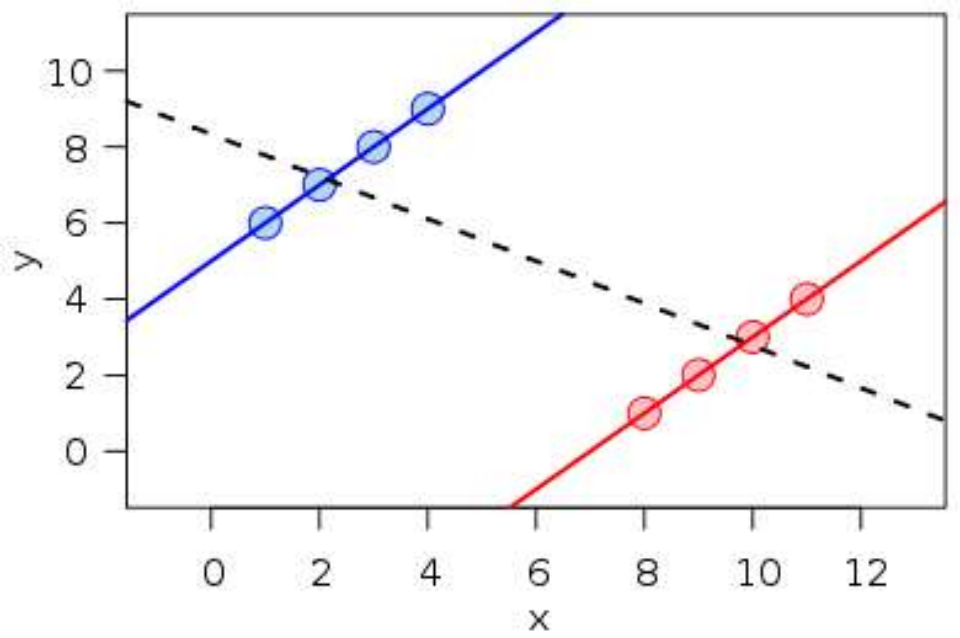
- Lazard et al., 2012
  - Discussion: Significant Effect of CI Brand
    - Results reflect situation between 2002 and 2011
      - Do not take into account technical improvements during that time
    - 14% difference between best and poorest device
      - 0-100 range existed in CI speech performance in quiet

- Lazard et al., 2012
  - Discussion: Significant Effect of CI Brand
    - CI strategy was not recorded – default strategy assumed
    - Performance of each brand may vary based on characteristic tested
    - Other studies found different results comparing brands of CI
      - Cites Spahr et al., 2007

Simpson's Paradox

- A trend that appears in different groups of data may disappear when these groups are combined

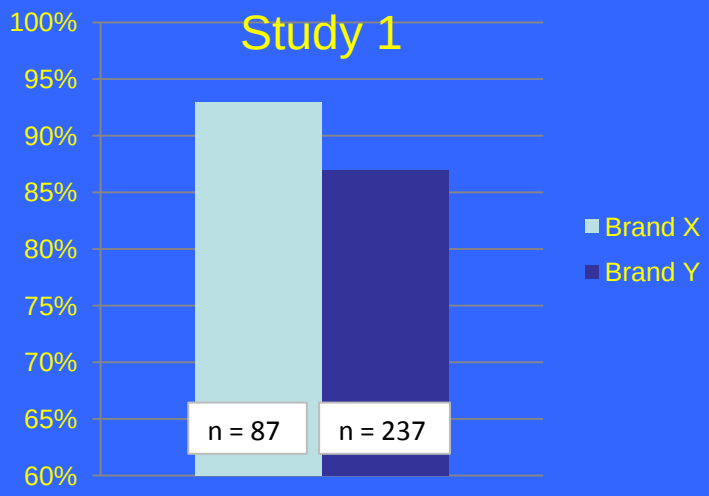
- A reverse trend appears for the aggregate data

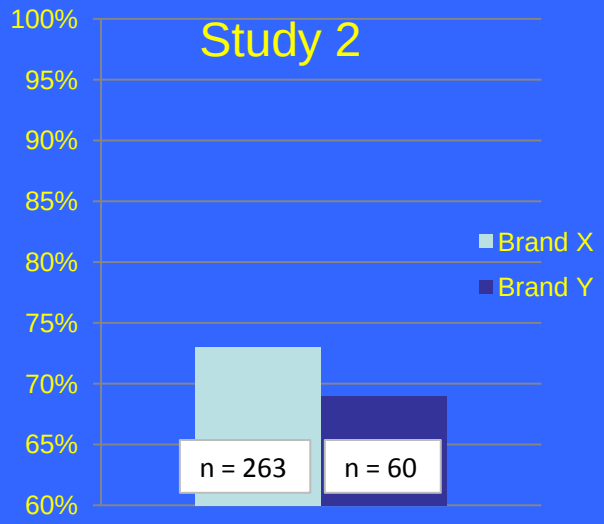


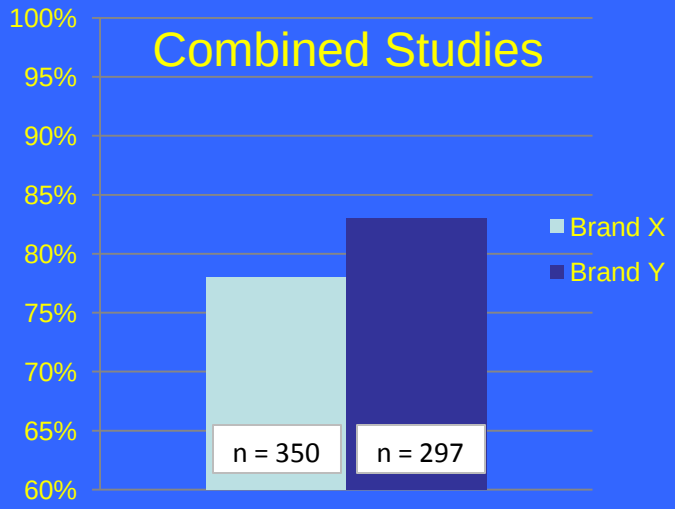
Simpson's Paradox:Example

- Consider a fictitious experiment in which 2 separate studies are conducted
  - Subjects randomly assigned to implantation with Brand X or Brand Y for each study
- Average speech recognition scores calculated for each device
- Then, both studies are compiled and the weighted average score is calculated for each device

Example of Simpson's paradox







Conclusion

- To date, two high profile studies looked like they might have shown the superiority of one brand of cochlear implant over another
- Upon closer examination, the analysis methods used in each study are not appropriate to determine whether one device is better than the other
- The only study of this type that remains valid is the 1992 VA Randomized Controlled Trial by Cohen and Waltzman that showed the superiority of multichannel over single channel cochlear implants