NEXTEEL Co. v. United States
Consol. Court No. 18-00083, Slip Op. 23-52 (CIT April 19, 2023)
Oil Country Tubular Goods from the Republic of Korea

FINAL RESULTS OF REDETERMINATION PURSUANT TO COURT REMAND

I. SUMMARY

The U.S. Department of Commerce (Commerce) has prepared these final results of redetermination pursuant to the opinion of the United States Court of International Trade (CIT) in *NEXTEEL v. United States*, Consol. Court No. 18-00083, Slip Op. 23-52 (CIT April 19, 2023) (*Remand Order*). This redetermination concerns the final results of the administrative review of the antidumping duty (AD) order on certain oil country tubular goods (OCTG) from the Republic of Korea (Korea) for the period of review (POR) September 1, 2015, through August 31, 2016.[1]

In the *Remand Order*, the CIT sustained Commerce's redetermination,[2] filed under respectful protest, that the alleged particular market situation did not exist during the POR in Korea.[3] Further, the CIT remanded Commerce's redetermination with respect to certain aspects of application of Cohen's *d* test in light of the opinion of the U.S. Court of Appeals for the Federal Circuit (CAFC) in *Stupp*.[4] In the *Remand Order*, the CIT found that Commerce's explanation in its *Third Redetermination* failed to resolve the concerns raised by the CAFC in

[1] *See Certain Oil Country Tubular Goods from the Republic of Korea: Final Results of Antidumping Duty Administrative Review; 2015-2016*, 83 FR 17146 (April 18, 2018) (*Final Results*), and accompanying Issues and Decision Memorandum (IDM).
[2] *See Final Results of Redetermination Pursuant to Court Remand, NEXTEEL CO. v. United States*, Consol. Court No. 18-00083, Slip Op. 21-1334 (Fed. Cir. March 11, 2022), dated October 21, 2022 (Third Redetermination), available at https://access.trade.gov/resources/remands/21-1334.pdf.
[3] *See Remand Order* at 7-18.
[4] *Id.* at 19 (citing *Stupp Corp. v. United States*, 5 F.4th 1341 (Fed. Cir. 2021) (*Stupp*)).

NEXTEEL IV that originated in *Stupp* relating to the use of the 0.8 threshold when certain statistical criteria (*i.e.*, normality, homoscedasticity, and sufficient number of observations) had not been addressed as part of Commerce's application of the Cohen's *d* test.[5] Additionally, in response to *NEXTEEL IV*, Commerce placed none of the academic literature discussed in *Stupp* on the administrative record of the previous remand segment, even though certain of these texts were used by SeAH Steel Corporation (SeAH) and Commerce in the IDM that accompanied the *Final Results* of the underlying administrative review. The CIT held that because Commerce relied on certain academic literature cited in SeAH's administrative case brief in its analysis supporting the *Final Results*, Commerce effectively made that academic literature part of the administrative record of this review. Accordingly, the CIT directed SeAH to place the academic literature included by Commerce in the *Final Results* IDM on the administrative record of this remand segment.[6] The CIT further remanded "this matter to Commerce for reconsideration of the academic literature cited in the {*Final Results*} IDM."[7]

For these final results of redetermination, and consistent with the *Remand Order*, we have confirmed that SeAH placed on the record of this remand segment the academic literature referenced by Commerce in the *Final Results* IDM. Additionally, consistent with the *Remand Order*, we have reconsidered this academic literature and analyzed the relationship of the 0.8 threshold with the statistical criteria and whether the statistical criteria are relevant to Commerce's application of the Cohen's *d* test. We conclude that the certain statistical criteria

[5] *See NEXTEEL Co. v. United States*, 28 F.4th 1226, 1239 (Fed. Cir. 2022) (*NEXTEEL IV*) ("SeAH argues Commerce's methodology was flawed because Commerce relied on Cohen's *d* even though the express conditions for its application were not satisfied: that the data sets being compared be normally distributed, have at least 20 or more data points, and have roughly equal variances." (internal citation omitted)).
[6] *See Remand Order* at 28.
[7] *Id.*

first addressed in *Stupp* need not be observed in the application of the Cohen's *d* test as part of Commerce's differential pricing analysis.

Finally, due to the exigencies of a burgeoning workload, including the unexpected filing of 27 new AD and countervailing duty petitions in the months of April and May 2023, Commerce was not able, within the time allotted, to prepare and issue draft results of redetermination to parties for comment, as is our normal and preferred practice.[8]

II. BACKGROUND

The statute provides that Commerce will compare normal value with U.S. price based on one of two standard comparison methodologies: the average-to-average method or the transaction-to-transaction method.[9] Alternatively, pursuant to section 777A(d)(1)(B) of the Tariff Act of 1930, as amended (the Act), Commerce may consider the use of the average-to-transaction method when two requirements have been met: (1) "there is a pattern of export prices (or constructed export prices) for comparable merchandise that differ significantly among purchasers, regions, or periods of time" (the pattern requirement);[10] and (2) "the administering authority explains why such differences cannot be taken into account using {the average-to-average method} or {the transaction-to-transaction method}" (the meaningful difference requirement).[11] Commerce's current practice is to use a "differential pricing analysis" to examine whether the two statutory requirements are satisfied.[12] As part of its differential pricing

[8] The filing of new AD/CVD petitions by outside parties is unpredictable, and the statutory initiation window of 20 days provides no flexibility. Significant agency resources must be allocated to accomplish this work in the extremely limited time permitted under the statute.

[9] *See* section 777A(d)(1)(A) of the Act and 19 CFR 351.414.

[10] *See* section 777A(d)(1)(B)(i) of the Act.

[11] *See* section 777A(d)(1)(B)(ii) of the Act.

[12] *See Certain Oil Country Tubular Goods from the Republic of Korea: Preliminary Results of Antidumping Duty Administrative Review; 2015-2016*, 82 FR 46963 (October 10, 2017) (*Preliminary Results*), and accompanying Preliminary Decision Memorandum (PDM) at 6-9; *see also Stupp*, 5 F.4th at 1346.

analysis, Commerce uses the Cohen's *d* test to determine whether prices differ significantly,[13] and the ratio test to determine whether the extent of significant price differences demonstrates a pattern of prices that differ significantly.

With regard to the Cohen's *d* test, the CIT remanded a narrow methodological issue for further explanation by Commerce. Specifically, the CIT found that Commerce's explanation in response to *NEXTEEL IV* "does not resolve the CAFC's concerns raised in *Stupp* pertaining to the use of the 0.8 threshold when the statistical assumptions are not observed."[14] Citing *Stupp*, the CIT noted that:

> Professor Cohen derived his interpretive cutoffs under certain assumptions. Violating those assumptions can subvert the usefulness of the interpretive cutoffs, transferring what might be a conservative cutoff into a meaningless comparison.[15]

The CIT concluded that it "remands for reconsideration or further discussion the issue of Commerce's calculation and application of the 0.8 threshold in {the} Cohen's *d* analysis."[16]

In the *Remand Order*, the CIT also noted that certain academic literature, although both discussed in SeAH's case brief and referenced by Commerce in the *Final Results*, was not on the record of this proceeding.[17] The CIT further noted that Commerce declined to consider this academic literature because it was not part of the administrative record of this proceeding.[18] The CIT acknowledged that "though the statistical limitations were drawn from academic literature,

[13] *See Preliminary Results* PDM at 9-10 ("The Cohen's *d* coefficient is a generally recognized statistical measure of the extent of the difference between the mean (*i.e.*, weighted-average price) of a test group and the mean (*i.e.*, weighted-average price) of a comparison group … . For this analysis, the difference is considered significant, and the sales in the test group are found to pass the Cohen's *d* test, if the calculated Cohen's *d* coefficient is equal to or exceeds the large (*i.e.*, 0.8) threshold.").

[14] *See Remand Order* at 23; *see* also *NEXTEEL IV*, 28 F.4th at 1239 ("Because Commerce's use of Cohen's *d* here presents identical concerns to those in *Stupp*, we vacate this portion of *NEXTEEL I* and remand to the Court of International Trade to reconsider in view of *Stupp*." (internal citations omitted)).

[15] *See Remand Order* at 23 (citing *Stupp*, 5 F.4th at 1360).

[16] *Id.*

[17] *Id.* at 23-24; *see also* SeAH's Letter, "Case Brief of SeAH Steel Corporation" dated November 30, 2017, at 25-42; and *Final Results* IDM at Comment 8.

[18] *See Remand Order* at 23-28.

{Commerce} was not required by the CAFC to incorporate the academic literature into its response."[19] Nonetheless, the CIT further determined that Commerce's inclusion of this academic literature in the IDM accompanying the *Final Results* "effectively made the academic literature part of the administrative record."[20] Accordingly, the CIT ordered SeAH to place on the administrative record the academic literature cited by Commerce in the *Final Results* IDM.[21] Between May 31 and June 12, 2023, SeAH placed on the record of this remand segment copies of the academic literature specified by the CIT in its *Remand Order*.[22]

Our analysis of the use of Dr. Cohen's large, 0.8, threshold when the statistical criteria need not be observed and the academic literature referenced by the CIT in the *Remand Order* is presented below.

III. ANALYSIS

In *Stupp*, the CAFC recognized that *Mid Continent 2019* had resolved the issue of whether Commerce's adoption of the large, 0.8, threshold was reasonable,[23] but noted that it did not reach the narrow question of whether the 0.8 threshold could be applied when the statistical assumptions (*i.e.*, normality of the distribution, equal variances, and roughly the same number of

[19] *Id.* at 25.

[20] *Id.* at 28.

[21] *Id.*

[22] *See* SeAH's Letter, "Resubmission of Publications Pursuant to Department's May 26 Letter," dated May 31, 2023 (includes Attachment 1 - Coe, Robert, "It's the Effect Size Stupid: What Effect Size Is and Why It Is Important," paper presented at the Annual Conference of the British Educational Research Association (September 2002) (*Coe*), and Attachment 4 - Lane, David, *et al.*, *Introduction to Statistics*, Online Edition, Chapter XIX, Part 3: "Difference Between Two Means" (*Lane*)); and SeAH's Letter, "Resubmission of Publications Pursuant to Department's June 8 Letter," dated June 12, 2023 (includes Attachment 1 - Ellis, Paul D., *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*, Cambridge University Press, 2010 (*Ellis*), and Attachment 2 - Cohen, Jacob, *Statistical Power Analysis for the Behavior Sciences*, Second Edition, Lawrence Erlbaum Associates (1988) (*Cohen*)) (collectively, Academic Literature).

[23] *See Stupp*, 5 F.4th at 1357 ("We held that 'the 0.8 standard is 'widely adopted' as part of a 'commonly used measure' of the difference relative to such overall price dispersion … . {I}t is reasonable to adopt that measure where there is no better, objective measure of effect size.'" *Mid Continent Steel & Wire, Inc. v. United States*, 940 F.3d 662, 673 (Fed. Cir. 2019) (*Mid Continent 2019*)).

observations (*i.e.*, the sample size))[24] identified by the CAFC[25] are not observed. In particular, the CAFC in *Stupp* expressed concerns about application of the 0.8 threshold when the statistical criteria had not been observed, stating that "Professor Cohen noted that 'we maintain the assumption that the populations being compared are normal and with equal variability, and conceive them further as equally numerous.'"[26] After reviewing the academic literature placed on the record, we do not find that the academic literature on the record provides justification to cause us to deviate from the use of the 0.8 threshold in the context of our differential pricing analysis or renders this widely-accepted threshold meaningless or unreasonable.

With the Academic Literature on the record pursuant to the *Remand Order*, Commerce has reexamined the origin of Dr. Cohen's thresholds, as well as the role of the statistical criteria. Based on our review of the academic literature, we find no support in the Academic Literature for the claim that Dr. Cohen's 0.8 threshold was derived based on the statistical criteria, or that the use of Dr. Cohen's threshold should be limited to situations where the sampled data exhibit a normal distribution or similarly equal variances. Dr. Cohen proposed his small, medium, and large thresholds as a convention; he expected that, while "arbitrary," they "will be found to be reasonable by reasonable people."[27] The actual numerical values for Dr. Cohen's proposed thresholds (*i.e.*, 0.2, 0.5, and 0.8 for small, medium, and large effects, respectively) were not based on research results or statistical analyses, but were threshold numbers that Dr. Cohen

[24] *See Stupp*, 5 F.4th at 1356 ("SeAH argues that Commerce's application of the 0.8 cutoff in this case was unsupported by evidence because Professor Cohen's suggestion that '0.8 could be considered a 'large' effect size' was limited to comparisons involving data that met certain restrictive conditions—'in particular, that the datasets being compared had roughly the same number of data points, were drawn from normal distributions, and had approximately equal variances.'" (internal citation omitted)).

[25] *See Stupp*, 5 F.4th at 1357.

[26] *Id.* (citing *Cohen* at 21; *and Cohen* at 25-26 ("discussing 'small effect size' 0.2, 'medium effect size' 0.5, and 'large effect size' 0.8 '{i}n terms of measures of nonoverlap … of the combined area covered by two normal equal-sized equally varying populations.'")

[27] *See Cohen* at 13.

proposed because he considered that they will be found reasonable by others.[28] Having reviewed the Academic Literature, we find no basis to conclude that the statistical criteria, which raised concerns before the CAFC in *Stupp*, were part of Dr. Cohen's selection of these proposed conventions. To the contrary, the academic literature states that "effect sizes exist in the real world" and "{t}he best way to measure an effect is to conduct a census of an entire population but this is seldom feasible in practice."[29]

The statistical assumptions are necessary to ensure that a selected sample properly represents the whole population; they are independent of the threshold established by Dr. Cohen. The effect size calculated from a properly drawn sample is reflective of the whole population and is accepted by Dr. Cohen and researchers as being confidently representative of an outcome as if the entire population were examined. Therefore, an effect size that is based on a whole population is equally representative of the population, and researchers are confident that the results are reflective of the whole population. Because a properly drawn sample and a whole population are both representative and reflective of the whole population, both can similarly be compared to the established thresholds.

In addressing the CIT's concern about not observing the assumptions when examining the whole population, it is important to understand why researchers examine the statistical

[28] *See Mid Continent 2019* ("{T}he 0.8 standard is 'widely adopted' as part of a 'commonly used measure' of the difference relative to such overall price dispersion; and it is reasonable to adopt that measure where there is no better, objective measure of effect size. We agree with the Trade Court that this rationale adequately supports Commerce's exercise of the wide discretion left to it under {section 777A(d)(1)(B) of the Act}." (citing *Certain Steel Nails from Taiwan: Final Determination of Sales at Less Than Fair Value*, 80 FR 28959 (May 20, 2015), and the accompanying IDM at 25-26 ("In 'Difference Between Two Means,' the author states that 'there is no objective answer' to the question of what constitutes a large effect. Although {respondent} focuses on this excerpt for the proposition that the 'guidelines are somewhat arbitrary,' the author also notes that the guidelines suggested by Cohen as to what constitutes a small effect size, medium effect size, and large effect size 'have been widely adopted.' The author further explains that Cohen's *d* is a 'commonly used measure{}' to 'consider the difference between means in standardized units.'" (quoting *Lane* at 1-2))))

[29] *See* Ellis at 5.

assumptions in the first place. All the Academic Literature and arguments presented focus on the context of sampling, and the statistical assumptions are assessed with the goal of improving the research results. The aim of inferential statistics is to have a certain confidence level or high probability that conclusions drawn from a sample accurately reflect what the results would be as if the entire population were studied. These assumptions are used to ensure that selected samples meet the established parameters (*e.g.*, normal distribution, equal variances, confidence level, margin of error, the sample size) required by the researchers. These variables are interconnected and help researchers achieve the required probability that the outcome derived from samples accurately reflects the entire population.

Understanding the relationship between selecting a sample from the population versus using the whole population, and the impact of the assumptions used, requires a fundamental understanding of the purpose behind selecting a sample, as well as the parameters examined when selecting a sample to ensure the results properly reflect an outcome as if the study were conducted on the whole population.

Because examining the whole population is typically not practical or not cost-effective, when conducting research, researchers must identify an appropriate sample size to test the hypothesis. Sampling is a statistical technique used to select a representative subset of observations from a population to participate in a study, which can be used to make inferences about the population as a whole. Sampling reduces the cost and time required for research studies, increases the accuracy and reliability of the obtained results when the population cannot be examined in its entirety, and is often the only practical way to deal with large populations.[30]

[30] *See, e.g.*, Ellis at 5.

While researchers are interested in an entire group, when studying a large population, it is generally not feasible to survey the entire population. Hence, researchers will take a random sample to approximate or represent the population as a whole. The size of the sample is important to achieve accurate, statistically significant results and to successfully run a study. If the sample is too small, researchers may include a disproportionate number of sampled observations that are outliers and anomalies from the population. These skew the results and do not provide researchers with a representative picture of the whole population. If the sample is too large, researchers may find the study to be complex, costly, time-consuming, and unrealistic to conduct. While the results will be more representative of the population, the benefits may not outweigh the costs. Therefore, one of a researcher's first steps when conducting a study is developing the parameters for selecting an appropriate sample from the whole population that meets the needs of the researchers. This is where the assumptions come into play. Once researchers know the size of the whole population, identify the confidence level desired, select an acceptable margin of error and desired dispersion, using the statistical formula for selecting a sample size, or simply using an online "sample size calculator"[31] can determine the exact number of observations necessary to achieve a high probability that the outcome represents the entire population.

In contrast, when researchers examine the whole population, concerns related to the parameters of distribution, variances, and number of observations are not relevant because the parameters of the whole population being examined are already established and properly reflect the characteristics of the whole population. Naturally, these statistical assumptions, which are necessary for a sample to accurately estimate/represent the whole population, are therefore, not

[31] *Id.* at 14.

needed. In fact, modifying the data in any way when using the whole population distorts the actual population's parameters, making it no less reflective of the whole population and, thereby, decreasing the confidence level of the data relied upon.

Based on our review, the cited literature focuses on the application of Cohen's *d* test in the context of research involving samples and does not contain any express mention of criteria or assumptions necessary when examining an entire population. These issues, which the assumptions attempt to address, arise only when researchers draw inferences about a population based on a sample. Inferential statistics are applied in an effort to make the samples as representative of the whole population as possible and to draw inferences from those results and project them to the population. This is the foundation and basic concept underpinning inferential statistics.[32] While the argument has been made that the assumptions applied to samples must also be applied when using the whole population, there is no logical explanation as to why such assumptions are necessary.

The purpose behind the three statistical criteria (*i.e.*, normality of the distribution, equal variances, and the size of the sample (*i.e.*, number of observations)) is to make samples more reflective of the population, which in turn increases the confidence level that the results are reflective of the whole population. If researchers examine the whole population, these assumptions become unnecessary, as there is no need to make a whole population more reflective of the population. To repeat, any adjustment to the population data serves only to distort the actual population's parameters, making it no longer reflective of the whole population and reducing the 100 percent confidence level. Nowhere in the cited literature is there any mention of criteria or assumptions necessary when examining the entire population. In contrast,

[32] *Id.* at, *e.g.*, 20.

Dr. Cohen's thresholds do not depend on the subjective composition of a particular sampled population. The only references to assumptions are related to drawing a sample and to efforts to improve the probability that the samples are as representative of the whole population as possible.

With the Academic Literature on the record pursuant to the *Remand Order*, Commerce reexamined the origin of Dr. Cohen's thresholds as well as the role of the statistical criteria. The analysis supports the position that Dr. Cohen's thresholds are widely accepted thresholds[33] and, thus, do not depend on subjective composition of a particular population (including number of observations, variance, and distribution). Commerce's application of the Cohen's *d* test and Dr. Cohen's thresholds to the entire population of relevant price observations does not require the application of the three statistical criteria identified by the CAFC in *Stupp*. Below, we present an illustrative framework to describe the relationship between Dr. Cohen's thresholds and the statistical criteria, and we identify the academic literature to support Commerce's position.

The purpose of the statistical criteria is to determine whether analysis results which are based on sampled data are representative of the results if the analysis had been based on the full population of data. For example, a COVID-19 vaccine is tested on 1,000 people to determine whether it will be safe and effective for the general population. The statistical criteria are part of the analysis to establish whether the results concerning safety and efficacy, which are found for the sample of 1,000 people, are representative of the safety and efficacy of the vaccine when given to millions of people. The role of the statistical criteria (*i.e.*, type(s) of distribution, variance(s) and sample size(s)) is to be part of the analysis to determine the "reliability of {the}

[33] *See Mid Continent 2019* ("It is reasonable to adopt that measure where there is no better, objective measure of effect size.'").

sample results."[34] Commerce's application of the Cohen's *d* test, including Dr. Cohen's large, 0.8, threshold, do not require addressing the three statistical criteria identified by the CAFC in *Stupp*. Dr. Cohen's large threshold is not dependent on the statistical criteria, and because the prices used in the Cohen's d test include all prices of comparable merchandise for the test and comparison groups (*i.e.*, akin to drawing conclusions of the efficacy of the vaccine from a study that encompasses the entire population), there is no role for the statistical criteria to examine whether the test results are reliable and representative of the results if calculated on the full populations of data. Below, we present an illustrative framework to further illustrate the relationship between Dr. Cohen's thresholds and the statistical criteria, and we identify the academic literature to support Commerce's position.

A. Illustrative Framework

The task is to determine whether BigBill's prices of bicycles differ significantly between Virginia and Maryland in 2020. All of the bicycles sold by BigBill in the two states are comparable merchandise. To determine whether BigBill's prices in Maryland and Virginia differ significantly, we will use the concept of effect size, and specifically Dr. Cohen's *d* coefficient, where the difference in the mean prices will be measured relative to the variance (*i.e.*, standard deviation) in the prices in each state. Further, we have decided to use Dr. Cohen's large, 0.8, threshold to determine whether the difference in prices in the two states is significant.

For our analysis, we randomly select five sale prices from each state, and calculate the mean and standard deviation of the five sale prices in each market. We also calculate a Cohen's *d* coefficient based on the sampled data, which by happenstance results in 0.9, *i.e.*, passing. However, because we examined only a small percentage of the actual number of sales made, we

[34] *See Cohen* at 6.

determine that the reliability of the sampled data, based in part on the statistical criteria, is not statistically significant. Consequently, we must revise the analysis.

For the revised analysis, we randomly select twenty sale prices from each state, and calculate the mean and standard deviation of the sampled prices from each state. We also calculate a Cohen's *d* coefficient, which now results in 0.75, a finding that the prices do not differ significantly between the two states. Finally, and just as importantly, we find based in part on the statistical criteria, that the calculated results using the sample reliably represent the results as if the calculations had been based on the full populations of sale prices to each state. Accordingly, we conclude, based on our analysis of sample sale prices in each state, that the difference in all of BigBill's prices of bicycles in Maryland and Virginia is not significant.

Next, there is a second analysis to determine the sale prices of exotic sport cars differ significantly between Vermont and New Hampshire. The cars sold in each state have prices that are comparable. During 2020, two cars were sold in Vermont and five were sold in New Hampshire. From the results of the previous analysis, we question whether we can do the analysis, because the previous analysis demonstrated that we needed a larger sample size (*e.g.*, twenty or more) so that the results calculated based on the sample data reliably represent the actual parameters of the full populations of data. However, if we make our calculations based on all of the sale prices within each state, then the calculated parameters (*i.e.*, the means, standard deviations and Cohen's *d* coefficient) will be the actual parameters of all sale prices to each state (*i.e.*, the full populations of data), even when the full populations are comprised of small numbers of observations, in this scenario, two and five. Unlike the example involving bicycle sale prices that have been sampled, the calculated results here that are based on sale prices of all cars will not change because of different samples of data. This is not dependent upon how many sales are made in each state (*i.e.*, the number of observations in the data). The calculated values of these parameters are not estimates of the actual values of those parameters; they are the actual values. Further, regarding the reliability of the results based on meeting certain statistical assumptions (*i.e.*, adequate number of observations), although the number of observations is small, the results reflect the actual values of the full populations of sale prices. Further, the reliability of those values based on the sales sampled to represent the population is not relevant.

In each analysis, we have used a measure of effect size, *i.e.*, Dr. Cohen's *d* coefficient, and Dr. Cohen's large, 0.8, threshold to decide whether the difference in the mean prices is significant. The effect size is a characteristic, *i.e.*, the degree that a phenomenon exists, that is inherent in the population. Dr. Cohen's small, medium, and large thresholds are one way to understand, to interpret, a measure of effect size, which is otherwise a unitless number.[35] Although effect size is a characteristic of a population, often the results of an analysis are based on sampled data and, thus, the various parameters (*e.g.*, mean, standard deviation, effect size) calculated based on sampled data are estimates of the actual values based on the full population of data. To demonstrate that analysis results are representative of the actual parameters of the population, statistical inference must be used based on characteristics of the sampled data, including the statistical criteria of concern before the CAFC. This is not the case, however, in Commerce's differential pricing analysis in which the calculations do not involve samples or assumptions; they are rather based on the entire population of comparable price observations and result in the actual parameters of the population.

[35] An example of a measure with units could be a dog's weight, where one canine that weighs 20 pounds is small, and another that weighs 100 pounds is large.

B. Academic Literature

With the above framework to provide illustration, the academic literature, especially Dr. Cohen's text, describes the effect size as a characteristic, or a phenomenon, of the population of data. For the above framework, the phenomenon is the difference in prices between two states for comparable merchandise. Moreover, the statistical criteria, along with the significance criterion, establish whether calculated results based on sampled data reliably represent (or estimate) the actual phenomenon in the population. Finally, Dr. Cohen's thresholds, which may be used to define whether prices differed significantly, are independent of statistical inferences and are simply numbers which have been widely accepted as one alternative to interpret a calculated effect size.

The purpose of Dr. Cohen's text is to guide researchers in the development of analyses based on sampled data, such that the results of those analyses will satisfy the pre-established assumptions of the study. Such development of a research project must balance the uncertainty of sampling with the resources that are available to the researcher. Dr. Cohen defines effect size as a component of such an analysis. In the context of such research analyses, Dr. Cohen presents effect size as a phenomenon of the population, which the research analysis will include.

Dr. Cohen's Statistical Power Analysis

Dr. Cohen's text, *Statistical Power Analysis for the Behavioral Sciences*, presents the concept of a "power analysis,"[36] which tests the null hypothesis to determine whether a phenomenon in a population exists based on a sample.[37] In the examples above, the

[36] *See Cohen* at 1 (Dr. Cohen's purpose is "to provide a self-contained comprehensive treatment of statistical power analysis from an 'applied' viewpoint" where the "power of a statistical test is the probability that it will yield statistically significant results.").

[37] *Id.* at 1 (In general, the result that is sought is based on a test of the null hypothesis, "*e.g.*, 'the hypothesis that the phenomenon to be demonstrated is in fact absent'" but whereas a researcher "typically hopes to 'reject' this hypothesis and thus 'prove' that the phenomenon in question is in fact present." (internal citation omitted)) and 4

"phenomenon" is the difference in prices between the two states, and the null hypothesis is that the difference in prices is equal to zero (*i.e.*, identical). Rejection of the null hypothesis would indicate that there is a non-zero difference in the prices between the two states.

A power analysis is dependent on three parameters: (1) the significance criterion;[38] (2) the reliability of the sampled data;[39] and (3) the effect size.[40] "The degree to which the phenomenon is present in the population" is measured by the effect size, where the greater the phenomenon exists *in the population*, the larger the effect size.[41] In the above examples, if the null hypothesis is rejected, then the result of the analysis is that the prices differ by some non-zero amount. The extent that the prices differ between the two states is measured by the effect size.

Statistical Inference

The first two parameters of the power analysis, the significance criterion and the reliability of the sample data, evaluate whether the results based on a sample reliably represent the phenomenon in the full population of data.[42] This "statistical inference" is dependent on the probability of rejecting a true null hypothesis (*i.e.*, Type I error), the sample size, and for the

("*The power of a statistical test of a null hypothesis is the probability that it will lead to the rejection of the null hypothesis*, *i.e.*, the probability that it will result in the conclusion that the phenomenon exists." (emphasis in the original)).

[38] *Id.* at 4 ("{T}he significance criterion represents the standard of proof that the phenomenon exists, or the risk of mistakenly rejecting the null hypothesis." "{I}t is the rate of rejecting a true null hypothesis," *e.g.*, a Type I error.)

[39] *Id.* at 6 ("The reliability (or precision) of a sample value is the closeness with which it can be expected to approximate the relevant population value. It is necessarily an estimated value in practice, since the population value is generally unknown. Depending upon the statistic in question, and the specific statistical model on which the test is based, reliability may or may not be directly dependent upon the unit of measurement, the population value, and the shape of the population distribution. However, it is *always* dependent upon the size of the sample." (emphasis in the original))

[40] *Id.* at 9-10 (the "effect size {means} 'the *degree* to which the phenomenon is present in the population,' or 'the degree to which the null hypothesis is false.'" (emphasis in the original)).

[41] *See Ellis* at 5 ("The best way to measure an effect is to conduct a census of an entire population but this is seldom feasible in practice.").

[42] *See Cohen* at 1-2 (One cannot ignore "the necessarily probabilistic character of *statistical inference*" and that the "{r}esults from a random sample drawn from a population will only approximate the characteristics of the population." (emphasis added)).

difference of the means analysis, the shape of the population distribution (*i.e.*, normality and homoscedasticity).[43] In the examples above involving bicycle prices, statistical inferences are used to determine whether the results of the analysis based on sampled prices in each state reliably represent the price differences for all prices in each state. As this is based on a difference in the means, the statistical inference is dependent upon the statistical criteria, *i.e.*, sample size, normality and homoscedasticity, and the significance criterion.

Effect Size

In Dr. Cohen's text:

> Each of the Chapters 2-10 will present in some detail the {effect size} index appropriate to the test to which the chapter is devoted. Each will be translated into alternative forms, the operational definitions of 'small,' 'medium,' and 'large' will be presented, and examples drawn from various fields will illustrate the test. This should serve to clarify the {effect size} index involved and make the methods and tables useful in research planning and appraisal.[44]

Specifically, as "seen in Chapter 2, the {effect size} index for *differences between population means* is standardized by division by the common within-population standard deviation (σ)."[45] Thus, Dr. Cohen's *d* coefficient is a standardized, unitless ratio of the difference in the means divided by some measure of the dispersion of the data,[46] all of which are to be reflective of the whole population.

Dr. Cohen's Thresholds

"To this point, the {effect size} has been considered quite abstractly as a parameter which can take on varying values (including zero in the null case). In any given statistical test, it must be indexed or measured in some defined unit appropriate to the data, test, and statistical model

[43] *Id.* at 19-20.
[44] *Id.* at 13-14.
[45] *Id.* at 11 (emphasis added).
[46] *Id.* at 21 ("Since both numerator and denominator are expressed in scale units, these 'cancel out,' and *d* is a pure number (here a ratio), freed of dependence upon any specific unit of measurement.").

employed."[47] Dr. Cohen prompts the researcher to respond to the question, "How large an effect do I expect *exists in the population*?"[48] "{The researcher} may initially find it difficult to answer the question even in general terms, *i.e.*, 'small' or 'large,' let alone in terms of the specific {effect size} index demanded."[49] The answer to such a question may depend upon resources available to the researcher. Alternatively, Dr. Cohen proposed "*as a convention*, {effect size} values to serve as operational definitions of the qualitative adjectives 'small,' 'medium,' and 'large.' This is an operation fraught with many dangers: The definitions are arbitrary, such qualitative concepts as 'large' are sometimes understood as absolute, sometimes as relative; and thus they run a risk of being misunderstood."[50] Nonetheless, Dr. Cohen emphasizes that "{a}lthough arbitrary, the proposed conventions will be found to be reasonable by reasonable people."[51]

For an analysis based on the difference of the means, Dr. Cohen proposed that numerical thresholds to define a small, medium, and large effect, *i.e.*, 0.2, 0.5, and 0.8 respectively.[52] As discussed above, these numerical thresholds are arbitrary, but Dr. Cohen expected that they would be found reasonable.[53] Indeed, these thresholds have been "widely accepted" as recognized in *Mid Continent 2019*, and "Cohen's cut-offs provide a good basis for interpreting effect size and for resolving disputes about the importance of one's results."[54] Further, the Academic Literature provides no evidence that the values themselves or their use are dependent

[47] *Id.* at 11.
[48] *Id*. at 12 (emphasis added) and 20-21.
[49] *Id.*
[50] *Id.* (emphasis in original).
[51] *Id.* at 13.
[52] *See Cohen* at 24-27.
[53] *See Ellis* at 41 ("Cohen's effect size classes have two selling points. First, they are easy to grasp. You just compare your numbers with his thresholds to get a ready-made interpretation of your result. Second, although they are arbitrary, they are sufficiently grounded in logic for Cohen to hope that his cut-offs 'will be found to be reasonable by reasonable people'" (internal citation omitted)).
[54] *Id.* at 40.

on statistical analysis or the application of the statistical criteria as argued by SeAH. Indeed, their usefulness is based on their acceptance within the academic community.

Nonetheless, Dr. Cohen provided real-life "operational definitions" to illustrate small, medium, or large effects. The first is an observational description, where, for example, a "medium effect size is conceived as one large enough to be visible to the naked eye."[55]

The second description is based on the concept of "percent nonoverlap," where Dr. Cohen posits:

> If we maintain the assumption that the populations being compared are normal and with equal variability, and conceive them further as equally numerous, it is possible to define measures of nonoverlap (U) associated with *d* which are intuitively compelling and meaningful.[56]

For the percent nonoverlap, Dr. Cohen conceives two bell curves where the difference in the means is the difference in the variable at the peak of each bell curve which is by definition the mean of a normal distribution. The area underneath each bell curve that is not also underneath the second bell curve is the percent nonoverlap. Dr. Cohen's requirement that each population be normally distributed, have equal variances, and be equally numerous is to permit the calculation of the area of the nonoverlap of the two bell curves. A normally distributed bell curve is defined by a specific probability function, which when the variance of the bell curve is known, allows for the calculation of the area underneath that curve. Likewise, when two bell curves are placed over one another, and both bell curves are normally distributed with equal variances, then the percent nonoverlap, just like the percent overlap (*i.e.*, the area common under both curves) can be calculated. However, the requirements of normality and homoscedasticity only apply to the ability to calculate the percent (*i.e.*, area) of nonoverlap as one approach to

[55] *See Cohen* at 26.
[56] *Id.* at 21-23.

illustrate different effect size values.[57] These limitations do not apply to Dr. Cohen's thresholds themselves, but only to the calculations which permit this example of interpreting different effect sizes.

Dr. Cohen's third operational definition of each threshold is to present different real-life examples where small, medium, and large effects have been found. These involve the differences in the IQs of various groups of people or the differences in the heights of various ages of teenage girls.[58] These illustrative examples also do not link Dr. Cohen's thresholds with the statistical criteria, as the 0.8 effect which has been observed is for the population of, for example, all Ph.D. holders and college freshmen. Certainly, when the data on the IQs of these two groups of people were collected, it was not collected from everyone who met those definitions, but it would have been collected from a selected sample from each group. The results of the analysis would have been calculated based on the sampled data from each group, and also, through statistical inferences, the representativeness of those results for the entire populations would have been determined. If the statistical analysis of the sample demonstrated that the sample-based results are representative of the population, then the sample-based results would be applied to the entire populations of Ph.D. holders and college freshmen. This use of statistical inference, however, is necessary to ensure that the sample is representative, but it was not part of Dr. Cohen's proposed small, medium, and large thresholds, which are numerical values that have been widely accepted in the academic community.

[57] *Id.* at 22 and Table 2.2.1 (which presents the percent nonoverlap for various values of the Cohen's d coefficient. For example, for $d = 0$, the percent nonoverlap is 0.0 percent, *i.e.*, the bell curves lie completely on top of each other. For $d = 0.8$, the percent nonoverlap is 47.4 percent, or, in other words, almost half of the area under each of the bell curves is not common to both distributions).

[58] For example, a large effect "is represented by the mean IQ difference estimated between holders of the Ph.D. degree and typical college freshmen, or between college graduates and persons with only a 50-50 chance of passing in an academic high school curriculum. These seem like grossly perceptible and, therefore, large differences, as does the mean difference in height between 13- and 18-year-old girls, which is of the same size ($d = 0.8$)." *Id.* at 27 (internal citation omitted).

In the above examples, Dr. Cohen's large, 0.8, threshold is used as the definition that prices differ significantly between the two states. For the analysis involving exotic sports cars, the calculated effect size is based on all sale prices in each state, and thus, the large threshold is applied to the results of the calculations based on the full populations of prices. For the analysis involving bicycle sale prices, the result of the analysis is based on sampled prices. For the analysis where only five prices were selected from each state, the result was found to not be representative of the full population, and therefore, the analysis would not be determinative for all sale prices in the two states. However, for the analysis based on twenty sale prices selected from each state, the results of the analysis based on the sampled sale prices was found to be representative of all sale prices in the two states, such that the effect size calculated based on the sampled sale prices is considered to be the effect size for all sale prices in the two states. Thus, the comparison of the calculated effect size with the large threshold is a comparison of the effect size of the full populations of sale prices. Therefore, although the statistical criteria may be used to determine whether the result of an analysis is representative of the full populations of data, it is not part of Dr. Cohen's proposed thresholds to qualify an effect as small, medium, or large.

C. Commerce's Cohen's *d* Test

As discussed above, we find that the Academic Literature provides no evidence that Dr. Cohen depended upon or incorporated the statistical criteria when he established his proposed small, medium, and large thresholds for effect size. Moreover, Commerce's analysis in the Cohen's *d* test is to determine whether prices differ significantly between the sales to a specific purchaser, region, or time period (*i.e.*, the test group) and all other comparable sales (*i.e.*, the comparison group). These sale prices include *all* of the sale prices which are also used to calculate each respondent's weighted-average dumping margin and represent the *full population* of sale prices to each test and comparison group. As stated repeatedly, Commerce does not apply a sampling methodology when developing the test and comparison groups; Commerce relies on the entire populations of sales observations in both groups. Accordingly, use of the statistical criteria to determine the statistical significance of the calculated results is not relevant for the Cohen's *d* test or the differential pricing analysis as a whole.

IV. FINAL RESULTS OF REDETERMINATION ON REMAND

As a result of these finals results of redetermination, and consistent with the *Remand Order*, we have reexamined Commerce's use of Dr. Cohen's thresholds and the relevance of the statistical criteria identified in *Stupp* in the light of the Academic Literature. We continue to find that the statistical criteria are not relevant and do not limit Commerce's application of Dr. Cohen's thresholds in Commerce's Cohen's *d* test as part of the differential pricing analysis. Accordingly, we have not revised the analysis to calculate the weighted-average dumping margins calculated in this administrative review.

6/27/2023

X Lisa W. Wang

Signed by: LISA WANG

Lisa W. Wang
Assistant Secretary
for Enforcement & Compliance