

United States Court of Appeals for the Federal Circuit

**MARMEN INC., MARMEN ENERGIE INC.,
MARMEN ENERGY CO.,**
Plaintiffs-Appellants

v.

**UNITED STATES, WIND TOWER TRADE
COALITION,**
Defendants-Appellees

2023-1877

Appeal from the United States Court of International
Trade in No. 1:20-cv-00169-JCG, Judge Jennifer Choe-
Groves.

Decided: April 22, 2025

JAY CHARLES CAMPBELL, White & Case LLP, Washing-
ton, DC, argued for plaintiffs-appellants. Also represented
by RON KENDLER, ALLISON KEPKAY.

ROBERT R. KIEPURA, Commercial Litigation Branch,
Civil Division, United States Department of Justice, Wash-
ington, DC, argued for defendant-appellee United States.
Also represented by REGINALD THOMAS BLADES, JR., BRIAN
M. BOYNTON, PATRICIA M. MCCARTHY; JESUS NIEVES
SAENZ, Office of the Chief Counsel for Trade Enforcement

and Compliance, United States Department of Commerce, Washington, DC.

MAUREEN E. THORSON, Wiley Rein, LLP, Washington, DC, argued for defendant-appellee Wind Tower Trade Coalition. Also represented by THEODORE PAUL BRACKEMYRE, TESSA V. CAPELOTO, ROBERT E. DEFRADESCO, III, LAURA EL-SABAOWI, DERICK HOLT, ELIZABETH S. LEE, ALAN H. PRICE, JOHN ALLEN RIGGINS.

Before PROST, TARANTO, and CHEN, *Circuit Judges*.

PROST, *Circuit Judge*.

Marmen Inc., Marmen Énergie Inc., and Marmen Energy Co. (collectively, “Marmen”) appeal the U.S. Court of International Trade’s (“CIT”) decision sustaining the U.S. Department of Commerce’s (“Commerce”) final determination that calculated a 4.94% dumping margin for utility-scale wind towers from Canada. *See Marmen Inc. v. United States*, 545 F. Supp. 3d 1305 (Ct. Int’l Trade 2021) (“*Marmen I*”); *Marmen Inc. v. United States*, 627 F. Supp. 3d 1312 (Ct. Int’l Trade 2023) (“*Marmen II*”). We vacate and remand for further proceedings consistent with this opinion.

BACKGROUND

Wind towers “are a component of utility scale wind turbine electrical power generating units used to convert the energy from wind to electrical energy.” J.A. 49. “Wind towers are tubular steel structures upon which the other major wind turbine components *i.e.*, rotor blades and nacelles are mounted.” J.A. 49. “Wind towers have three types of sections: the base section, one or more mid-sections, and the top section.” Marmen’s Br. 5 (citing J.A. 74). “[S]teel plate is the primary input for wind towers,” and “the quantity and thickness of steel plate consumed in a particular tower can vary significantly depending on the specification.” J.A. 78.

In July 2019, Commerce initiated an antidumping (“AD”) investigation of wind towers from Canada in response to an AD petition filed by the Wind Tower Trade Coalition (“WTTC”). Marmen Inc. and Marmen Énergie Inc. were selected as mandatory respondents. In February 2020, Commerce issued a preliminary AD determination, assigning Marmen a 5.04% dumping margin, and in June 2020, Commerce issued its final AD determination, assigning Marmen a 4.94% dumping margin.

Marmen appealed to the CIT, making three arguments that are relevant here: Commerce erred by (1) determining that Marmen’s steel plate costs did not reasonably reflect the costs associated with the production and sale of the products, and thus it was error for Commerce to weight-average (or smooth) the reported steel plate costs; (2) rejecting a supplemental cost-reconciliation item that was intended to correct certain purchase information that had not been properly converted from U.S. dollars (“USD”) to Canadian dollars (“CAD”); and (3) using the average-to-transaction (“A-to-T”) methodology, instead of the ordinary average-to-average (“A-to-A”) methodology, to calculate Marmen’s dumping margin based on a misapplication of Cohen’s *d* test.

The CIT affirmed the weight-averaging of Marmen’s steel plate costs, *Marmen I*, 545 F. Supp. 3d at 1315, but remanded the case to Commerce on the other two issues. For the USD-to-CAD cost reconciliation, Commerce initially rejected the submission because it “constituted untimely and unsolicited new information.” *Id.* The CIT explained that “Commerce has a duty ‘to determine dumping margins as accurately as possible,’” *id.* at 1316 (quoting *NTN Bearing Corp. v. United States*, 74 F.3d 1204, 1208 (Fed. Cir. 1995)), and concluded that it was an abuse of discretion to deny the currency-conversion correction solely on the basis that it was untimely, *id.* at 1316–17.

As to the appropriate A-to-A or A-to-T methodology, Commerce first decides whether “a pattern of export prices (or constructed export prices) for comparable merchandise . . . differ significantly among purchasers, regions, or periods of time.” *Id.* at 1318 (quoting 19 U.S.C. § 1677f-1(d)(1)(B)(i)). If such a pattern exists, then Commerce “explains why such differences cannot be taken into account using [the A-to-A method].” § 1677f-1(d)(1)(B)(ii). If the A-to-A methodology cannot account for these differences, then Commerce may apply the A-to-T methodology instead to calculate a dumping margin. In *Marmen I*, Marmen argued that there was no significant difference among purchasers, regions, or periods of time, and thus Commerce should have used the A-to-A method to calculate the dumping margin. According to Marmen, Commerce erroneously found a significant price difference based on a misapplication of Cohen’s *d* test—a statistical test which can be used in certain circumstances to determine differences between the mean averages of two data sets. This, according to Marmen, led Commerce to incorrectly use the A-to-T methodology for calculating the dumping margin. *Marmen I*, 545 F. Supp. 3d at 1319. While *Marmen I* was pending before the CIT, our court issued *Stupp Corp. v. United States*, which called into question the appropriateness of Commerce’s use of Cohen’s *d* test. 5 F.4th 1341, 1357 (Fed. Cir. 2021) (“[T]here are significant concerns relating to Commerce’s application of the Cohen’s *d* test . . . in adjudications in which the data groups being compared are small, are not normally distributed, and have disparate variances.”). Based on *Stupp*, the CIT remanded “the issue of Commerce’s use of the Cohen’s *d* test for Commerce to explain further whether the limits on the use of the Cohen’s *d* test were satisfied.” *Marmen I*, 545 F. Supp. 3d at 1320.

On remand from *Marmen I*, Commerce again rejected the USD-to-CAD cost reconciliation because it allegedly would double count an exchange-rate adjustment already reflected in the cost of goods sold and reported cost of

production. *See Marmen II*, 627 F. Supp. 3d at 1316–20. As to Commerce’s application of Cohen’s *d* test, and despite concerns with its use expressed by our court in *Stupp*, Commerce determined that “the assumptions of normality and roughly equivalent variances’ are not relevant to Commerce’s application of the Cohen’s *d* test” because Commerce’s use of the test is based on an entire population of data and not a sample. *See id.* at 1320–22. The CIT, in *Marmen II*, agreed with Commerce on both issues and sustained Commerce’s determination of a 4.94% dumping margin.

Marmen timely appeals. We have jurisdiction under 28 U.S.C. § 1295(a)(5).

DISCUSSION

“While we recognize the [CIT] has unique and specialized expertise in trade law, its decision is reviewed *de novo*, applying anew the same standard used by that court in its consideration of Commerce’s determination.” *Union Steel v. United States*, 713 F.3d 1101, 1106 (Fed. Cir. 2013). “Accordingly, Commerce’s antidumping determination will be upheld unless it is ‘unsupported by substantial evidence on the record, or otherwise not in accordance with law.’” *Id.* (quoting 19 U.S.C. § 1516a(b)(1)(B)(i)). Substantial evidence is “such relevant evidence as a reasonable mind might accept as adequate to support a conclusion.” *Consol. Edison Co. v. NLRB*, 305 U.S. 197, 229 (1938). “[T]he relevant standard for reviewing Commerce’s selection of statistical tests and numerical cutoffs is reasonableness, not substantial evidence.” *Stupp*, 5 F.4th at 1353.

Marmen presents three main arguments on appeal: that (1) Commerce’s determination to weight-average (or smooth) Marmen’s steel plate costs was not supported by substantial evidence or otherwise in accordance with the law; (2) Commerce erred by rejecting Marmen’s USD-to-CAD cost reconciliation; and (3) Commerce’s use of the A-

to-T methodology based on Cohen's d test was unreasonable. We address each argument in turn.

I

Before calculating a dumping margin, Commerce must “determin[e] . . . whether subject merchandise is being, or is likely to be, sold at less than fair value.” 19 U.S.C. § 1677b(a). To do so, a comparison is made “between the export price or constructed export price and normal value,” *id.*, where the export price or constructed export price is the price at which subject merchandise is sold in the U.S. market and normal value is the price of a “foreign like product” sold in the producer’s home market or in a comparable third-country market. “If the price of an item in the home market (normal value) is higher than the price for the same item in the United States (export price), then the comparison produces a positive number, indicating that dumping has occurred.” *Thai Plastic Bags Indus. Co. v. United States*, 746 F.3d 1358, 1360 (Fed. Cir. 2014) (citing 19 U.S.C. § 1677(35)(A)). If a “foreign like product” is not identical to the merchandise sold in the U.S. market, then a “model match” methodology is used to sort merchandise into groups to be compared based on a hierarchy of similar characteristics. *See* § 1677(16). Each group is assigned a control number (“CONNUM”).

Here, Commerce compared the price of Marmen’s wind towers sold in the U.S. with those sold in its home market of Canada. There were no identical products, so the merchandise was sorted into groups based on similar characteristics. “In making product comparisons, [Commerce] matched foreign like products based on the physical characteristics reported by [Marmen] in the following order of importance: type (tower or section), weight of tower/section, height of tower/section, total sections, type of paint or coating, metalizing, electrical conduit – bus bars, electrical conduit – power cable, elevators, number of platforms, and other internal components.” J.A. 2466; *see also Marmen I*,

545 F. Supp. 3d at 1313. “Commerce determined that the most significant physical characteristics in differentiating costs of steel plate were type, thickness, weight, width, and height.” *Marmen I*, 545 F. Supp. 3d at 1314.

Once sorted into CONNUMs, the export price and the normal value may be compared. Under § 1677b, a comparison of the subject merchandise is based in part on the exporter’s cost of production. “Costs shall normally be calculated based on the records of the exporter or producer of the merchandise, if such records are kept in accordance with the generally accepted accounting principles of the exporting country (or the producing country, where appropriate) *and reasonably reflect the costs associated with the production and sale of the merchandise.*” § 1677b(f)(1)(A) (emphasis added). But when the reported costs do not reflect the production costs, then “it is Commerce’s normal practice to adjust costs to address distortions when such cost differences are attributable to factors beyond differences in the physical characteristics of such products.” J.A. 3858; *Thai Plastic Bags*, 746 F.3d at 1368 (“It is customary for Commerce to adjust a company’s reported allocation methodology to reflect costs based solely on physical characteristics.” (cleaned up)). In such circumstances, Commerce has “averaged” or “smoothed” certain CONNUM-specific costs across multiple CONNUMs.

That is what happened here—Commerce determined that the reported steel costs did not reasonably reflect the cost of production because “Marmen’s suppliers did not charge different prices for plates of different grade, thickness, width, or length,” and therefore, “there should be little difference in plate costs for different dimensions and grade based on record evidence on a per-unit weight basis.” *Marmen I*, 545 F. Supp. 3d at 1314. But Commerce did find differences unrelated to physical characteristics, such as the timing of production. *Id.* As a result, Commerce “weight-averaged the reported steel plate costs for all reported CONNUMs, except the CONNUM for the thickest

plate.” *Id.*¹ Marmen argues that this weight-averaging (or cost-smoothing) was improper. We disagree.

A

Marmen first argues that Commerce arbitrarily disregarded its standard test for when weight-averaging may be used—“Commerce’s consistent test for determining whether to average a respondent’s reported product-specific costs across multiple CONNUMs was to examine: (1) whether the respondent reported significantly different costs for ‘nearly identical’ or ‘similar’ CONNUMs and, if so, (2) whether such differences in cost were unrelated to the products’ physical characteristics.” Marmen’s Br. 8; *see also id.* at 34–35. Commerce responds that its “cost-smoothing practice is not limited solely to instances in which the products are ‘identical’ or ‘very similar.’” U.S. Br. 19. Instead, according to Commerce, “the key factor is whether a respondent’s reported cost differences properly reflect production differences associated with physical characteristics, as distinguished from other unrelated factors.” *Id.* The difference between these two positions, as relevant here, is whether Commerce may apply cost-smoothing practices only when the finished product (i.e., the subject merchandise) is nearly identical or similar, *or* whether Commerce may apply cost-smoothing practices when the input (e.g., raw materials) to the subject merchandise is nearly identical or similar. *See* Marmen’s Br. 34–35; U.S. Br. 19.

Commerce did not arbitrarily disregard its standard practice. Commerce’s statutory requirement is to

¹ “Commerce excluded the CONNUM for the thickest plates because the record indicated that there was a surcharge applied to high thickness plates that was not applied to lower thickness plates.” *Marmen I*, 545 F. Supp. 3d at 1314.

determine if the cost reporting “reasonably reflect[s] the costs associated with the production and sale of the merchandise.” § 1677b(f)(1)(A). Commerce has historically made that determination based on finished products in some cases and individual inputs in other cases. Indeed, in another antidumping case involving wind towers (described by Marmen as a companion case, Marmen’s Reply Br. 4 n.1), Commerce found that the per-unit weight cost of the steel plate input was “virtually the same regardless of the grade, thickness, width, or height.” *See Dongkuk S&C Co. v. United States*, No. 23-1419, slip op. at 12 (Fed. Cir. Apr. 21, 2025) (citation omitted). “Accordingly, Commerce concluded that ‘the overwhelming factor that caused the differences in the steel plate costs [for the final CONNUMs] was the timing of the steel plate purchases.’” *Id.* (alteration in original) (citations omitted). Based on these facts and findings (which are nearly identical to those before us), Commerce smoothed the respondent’s steel plate input costs. *Id.* at 7. We affirmed, concluding that “Commerce’s analysis did not have to focus on comparing the costs of finished wind towers sharing the same physical characteristics,” *id.* at 14, and “Commerce was not required to rely upon . . . distortive records and had the authority to adjust steel plate input costs to more accurately approximate [respondent’s] costs of production during the period of investigation,” *id.* at 11.

The same is true here. “It was reasonable for Commerce to analyze whether or not [respondent’s] costs were attributable to some relationship between raw material inputs and CONNUM physical characteristics.” *Id.* at 13. Therefore, as we did in *Dongkuk*, we conclude that Commerce did not act arbitrarily or inconsistently with its standard practice.

B

Marmen also argues that Commerce’s factual findings are not supported by substantial evidence. Specifically,

Marmen argues that “Commerce wrongly concluded that [Marmen’s] steel suppliers do not charge different prices for plates of different grade, thickness, width, or length, except for high thickness range plates (i.e., greater than 50.8mm),” and “Commerce assumed, contrary to the record evidence and its own findings, that differences in timing explained observed differences in Marmen’s reported CONNUM-specific plate costs.” Marmen’s Br. 36 (cleaned up). We agree with the CIT that Commerce’s factual findings are supported by substantial evidence.

As to Commerce’s determination that Marmen’s suppliers do not charge different prices for plates of different grade, thickness, width, or length, except for high thickness range plates, it was Marmen who first argued that “[t]hickness, in particular, affects steel plate cost.” J.A. 3709. Marmen also stated that multiple suppliers required upcharges for plates of the greatest thickness. J.A. 3709–10. Commerce, therefore, “excluded the CONNUM for the thickest plates because the record indicated that there was a surcharge applied to high thickness plates that was not applied to lower thickness plates.” *Marmen I*, 545 F. Supp. 3d at 1314. Marmen did not identify surcharges for the smaller thicknesses nor did Marmen identify surcharges or provide evidence of cost differences based on other physical characteristics. J.A. 3709–10. For example, despite stating that steel grade could cause differences in costs, Marmen confirmed that the grades from its suppliers were “roughly equivalent.” J.A. 3710. Based on Marmen’s own evidence, Commerce explained that “[u]nder Marmen’s theory, one would expect the thicker plates . . . to cost more per ton, and those that are thinner to cost less per ton.” U.S. Br. 26. But this was not the case. *See* Marmen’s Br. 37; U.S. Br. 26. Indeed, when Commerce compared the plate costs on a *per-unit weight basis*, it found a significant variance in the per-unit plate costs.

Having found that thickness, grade, and weight did not solely drive the cost, Commerce looked to other factors.

And in further analyzing the data, “Commerce found a pattern indicating that most of the CONNUMs with higher plate costs were sold early in the investigation period, whereas those with lower plate costs were sold later.” U.S. Br. 28 (citing J.A. 3874, 3879–80). On this record, we agree with the CIT that Commerce’s findings are supported by substantial evidence.

II

Next, Marmen argues that substantial evidence does not support Commerce’s decision to reject a minor correction to the cost-reconciliation worksheet to account for an exchange rate. On this question, we agree with Marmen.

The period of interest for the investigation was July 1, 2018, to June 30, 2019. During 2018, Marmen did not account for the exchange rate of USD-to-CAD for its “USD-denominated purchases.” Marmen’s Br. 12. During 2019, however, Marmen did convert its USD purchases to CAD in its financial records. Apparently relying on these financial records, when Marmen initially reported its costs of production in response to Section D of the AD Questionnaire, it inadvertently omitted the conversion from USD-to-CAD in the cost-reconciliation worksheet for Item L (“Affiliated purchase of wind sections from Marmen Énergie”) for the first half of the period of interest, i.e., July to December 2018. To correct this error, Marmen added a new line, Item L1, to adjust for the exchange rate from USD-to-CAD and submitted a Second Supplemental D Questionnaire on February 7, 2020. J.A. 3641. On February 25, 2020, Commerce rejected the correction as “untimely filed new factual information.” J.A. 3706–07; *see also* J.A. 3861.

In *Marmen I*, the CIT held that “Commerce abused its discretion by failing to consider Marmen’s corrective submission.” 545 F. Supp. 3d at 1317. On remand, however, Commerce again rejected the exchange-rate correction—this time alleging that the correction “would duplicate an adjustment amount that was already reflected in its

revised audited financial statements.” J.A. 4827. Marmen again appealed, and this time, the CIT accepted Commerce’s rejection of the exchange-rate conversion as supported by substantial evidence. On appeal here, Commerce makes two main arguments for why it rejected the exchange rate: (1) the exchange-rate conversion would be duplicative of other adjustments, U.S. Br. 40–41, and (2) Marmen’s specific exchange rate is unreliable, *id.* at 46. Neither argument has merit.

A

Commerce’s double-counting argument is unpersuasive. In a generous read of Commerce’s argument, its double-counting argument appears to be based on representations from Marmen’s auditors that the cost reconciliation was complete and allegations that line items P, Q, and R from Marmen’s cost-reconciliation worksheet already accounted for the exchange-rate conversion. But the statements from Marmen’s auditors were primarily made prior to the discovery of the omitted exchange-rate conversion. *See* U.S. Br. 40 (citing J.A. 1006–07; J.A. 4855–58 (almost entirely relying on evidence from before Marmen submitted the supplemental questionnaire)).

Equally unpersuasive is Commerce’s argument that “the cost reconciliation Item L, which is the figure that Marmen claims requires adjustment to account for U.S. dollar purchases from its French-Canadian affiliate, did not change based on the auditor’s adjustment to Marmen’s financial statement, indicating that the auditor did not believe any correction to that figure was necessary.” U.S. Br. 41 (citing J.A. 3905). We disagree. In Marmen’s February 7, 2020, submission, Marmen added Item L1, labeled “Exchange Rate Variance on July to Dec 2018 Affiliated Purchases of Wind Sections from Energie,” to adjust for the USD-to-CAD exchange rate. J.A. 3641 (“Version Two”). On February 25, Commerce rejected this submission as untimely and directed Marmen to “remove the entire” Item

L1. J.A. 3707. On February 28, 2020, Marmen complied and removed the line item. J.A. 3753 (“Version Three”). After *Marmen I*, Marmen submitted “Version Four” of the cost-reconciliation worksheet, again including Item L1 to adjust for the exchange rate. See J.A. 3904. In none of these versions did Marmen change Item L. Instead, it made the adjustment by adding the new Item L1. So the fact that Item L did not change does not indicate that Marmen did not make adjustments in the cost-reconciliation worksheet for Item L; Marmen just made the adjustment in a separate line item and labeled it accordingly. See J.A. 3904.

Commerce’s next argument is also unconvincing. Turning to line items P, Q, and R, Commerce admits that each of these line items are for exchange-rate conversions covering the first half of 2018. See J.A. 3905. We fail to see how adjustments for the first half of 2018 have any relevance to the proposed adjustments for the second half of 2018. Nor is there any explanation of how these exchange rates could be duplicative if they do not cover the same periods in time.

B

As to the alleged lack of reliability, Commerce first argues that it is unclear what transactions actually occurred in USD and which occurred in CAD. Commerce explains that when it reviewed the underlying invoices for the sales of the entire period of interest “almost every invoice listed in the document . . . is designated as a USD-denominated sale.” U.S. Br. 42 (citing J.A. 4858–59, J.A. 3907–13). But it then points to Marmen’s statements that the January to June 2019 purchases from Marmen Énergie were in CAD but not designated as such in the document listing the invoices. *Id.* Commerce’s argument appears to be that it cannot tell which sales are in USD and which are in CAD with any reliability. But this is an unnecessarily complicated view of the record. The records that Marmen relied on were

intended to show that the 2018 sales were recorded in USD and should be converted to CAD. That is what these records show. *See* J.A. 3907–09. Any other interpretation of the data amounts to a question of the veracity of Marmen’s representations—not its reliability. Commerce has not made an allegation that Marmen misrepresented its financial records nor are we aware of any records to support such a claim.

WTTC (the party who originally requested the AD investigation and also an appellee here) makes another attempt at a reliability argument—i.e., that Marmen provided no support for the average exchange rate for the relevant period. WTTC Br. 40–44. For example, WTTC argues that Marmen has confused which exchange-rate period should actually apply by pointing to Marmen’s assertion that the disputed exchange rate should apply for the second half of 2018 but then also stating that the exchange rate was “based on its exchange rate contracts in place *during the POI* [period of investigation].” WTTC Br. 40 (emphasis in original) (quoting Marmen’s Br. 15). WTTC’s reliance on this statement to confuse the period for the exchange rate is not consistent with the record as a whole. Marmen has been clear that the exchange rate it seeks to apply is for the second half of 2018, both in its record evidence and briefs. *See* Marmen’s Br. 12–18; J.A. 3904; J.A. 3907–09. Marmen’s statement that the requested exchange rate is based on exchange rate contracts in place *during the POI* does not undermine this, particularly because the second half of 2018 was *during the POI*.

Commerce and WTTC present additional arguments regarding the reliability of the proposed exchange rate, but we find these arguments equally unpersuasive. Thus, we conclude that Commerce’s determination that the exchange rate would double count records or is unreliable is not supported by substantial evidence. On this record, substantial evidence does not support Commerce’s rejection of the exchange rate.

III

Finally, we turn to Commerce's use of Cohen's *d* test. We begin with a discussion of how Cohen's *d* test is used in the process of calculating a dumping margin, and then turn to why normal distributions, equal variability, and equally and sufficiently numerous data are necessary to achieving a meaningful Cohen's *d* coefficient.

A

Our recent opinion in *Stupp* provides a thorough explanation of Cohen's *d* test and its relevance to calculating a dumping margin. 5 F.4th at 1344–48, 1357–60. The same statutory and regulatory background applies here. Therefore, we do not repeat all of this background information and instead only include a summary of Cohen's *d* test for the purposes of evaluating the arguments raised here. Additionally, *Stupp* involved a similar question regarding the appropriateness of using Cohen's *d* test in the context of calculating a dumping margin. The summary below includes the relevant conclusions from *Stupp* about Cohen's *d* test that led us to Commerce's arguments raised in the present case.

“When calculating a weighted average dumping margin, Commerce typically uses the average-to-average comparison method. 19 C.F.R. § 351.414(c)(1); *see also* 19 U.S.C. § 1677f-1(d)(1).” *Stupp*, 5 F.4th at 1345. But when “(i) there is a pattern of export prices (or constructed export prices) for comparable merchandise that differ significantly among purchasers, regions, or periods of time, and (ii) [Commerce] explains why such differences cannot be taken into account using [the A-to-A method],” § 1677f-1(d)(1)(B), Commerce may then use the A-to-T method. *Stupp*, 5 F.4th at 1345; *see also* U.S. Br. 48–49.

Cohen's *d* test is used as the first step in the differential pricing analysis to calculate the dumping margin.² First, Cohen's *d* test is used "to measure whether the United States prices to a particular purchaser, region, or time period differ significantly from the prices for all other purchasers, regions, and time periods." U.S. Br. 49–50 (citing *Stupp*, 5 F.4th at 1346). "If the Cohen's *d* value is equal to or greater than 0.8 for any test group, the observations within that group are said to have 'passed' the Cohen's *d* test, i.e., Commerce deems the sales prices in the test group to be significantly different from the sales prices in the comparison group." *Stupp*, 5 F.4th at 1347. "Commerce applies Cohen's *d* test to each test group within the regional, purchaser, and time-period categories." *Id.*

"Second, the ratio test calculates the proportion of respondent's United States sales, by value, that 'pass' the Cohen's *d* test, to determine whether a 'pattern' exists." U.S. Br. 50 (citing *Stupp*, 5 F.4th at 1347). Together, steps one and two determine, in accordance with the statute, whether "a pattern of export prices (or constructed export prices) for comparable merchandise that differ significantly among purchasers, regions, or periods of time." § 1677f-1(d)(1)(B)(i). Based on the outcome of these tests, Commerce selects which dumping margin calculation to use—either the A-to-A method or the A-to-T method (or some hybrid of the two).³ *See* U.S. Br. 47–52.

² "The differential pricing analysis involves three tests to address the statutory requirements of section 1677f-1(d)(1)(B)," U.S. Br. 49: (1) Cohen's *d* test, (2) the ratio test, and (3) the meaningful difference test. *See id.* at 49–51; *Stupp*, 5 F.4th at 1346–48.

³ The selection of the A-to-A method or A-to-T method is also subject to the meaningful difference test, but for the sake of simplicity, we do not discuss that test in

In *Stupp*, the appellant argued that “Commerce misused the Cohen’s *d* test in its differential pricing analysis.” 5 F.4th at 1357. We agreed with the appellant “that there are significant concerns relating to Commerce’s application of the Cohen’s *d* test in [that] case, and more generally, in adjudications in which the data groups being compared are small, are not normally distributed, and have disparate variances.” *Id.* “Our first concern is a general one: Commerce’s application of the Cohen’s *d* test to data that do not satisfy the assumptions on which the test is based may undermine the usefulness of the interpretive cutoffs.” *Id.* We explained that “[t]here is extensive literature describing the problems associated with applying the Cohen’s *d* test to data that are not normally distributed or that are lacking equal variances.” *Id.* at 1358; *see also id.* at 1358–59 (reviewing literature that investigated Cohen’s *d* when the assumptions are not met); *id.* at 1358 (applying Cohen’s *d* to non-normally distributed data sets and concluding “that applying Cohen’s *d* to such data caused serious flaws in interpreting the resulting parameter” (emphasis added)); *id.* (“conclud[ing] that Cohen’s *d* ‘was found to be inaccurate when the normality and homogeneity-of-variances assumptions were violated in this study, thereby severely affecting the accuracy of *d* in evaluating the true [effect size] in the research literature” (emphasis added) (second alteration in original)). Another “concern arises from test groups containing sales prices that hover around the same value” because it “tend[s] to artificially inflate the dumping margins for a set of export sales prices that has minimal variance.”

this opinion. *See* U.S. Br. 50; *Stupp*, 5 F.4th at 1347. Commerce did not use a meaningful difference test in a way that was independent of the first two steps, both of which depend on the use of Cohen’s *d* test. If the use of Cohen’s *d* test at the first and second steps is unsupported here, then we have no basis to find that error was harmless by use of the meaningful difference test. *See also infra* at 20.

Id. at 1359 (explaining mathematically how this artificial inflation of Cohen’s d coefficient may occur).

We concluded, “[i]n sum, the evidence and arguments before us call into question whether Commerce’s application of the Cohen’s d test to the data in this case violated the assumptions of normality, sufficient observation size, and roughly equal variances associated with that test.” *Id.* at 1360.⁴ But we did not reverse; rather we remanded the case to Commerce giving it an opportunity to address the following question:

We therefore remand to give Commerce an opportunity to explain whether the limits on the use of the Cohen’s d test prescribed by Professor Cohen and other authorities were satisfied in this case or whether those limits need not be observed when Commerce uses the Cohen’s d test in less-than-fair-value adjudications. In that regard, we invite Commerce to clarify its argument that having the entire universe of data rather than a sample makes it

⁴ *Stupp* addressed concerns with all three of Cohen’s d test’s required assumptions—normal distributions, equal variability, and equally and sufficiently numerous data—and concluded that the academic literature surrounding Cohen’s d test did not support the use of Cohen’s d test when these assumptions are not met. 5 F.4th at 1357–60. We do not repeat all of the reasons for the serious concerns with the use of Cohen’s d when these assumptions are not met but reaffirm those concerns are equally applicable to the facts before us. The opinion here instead focuses primarily on the question of whether using a population instead of a sample can assuage these concerns. (It does not.) We demonstrate this conclusion by focusing on the assumption of normal distribution because Commerce’s justification for using Cohen’s d test is largely based on reasoning regarding that assumption.

permissible to disregard the otherwise-applicable limitations on the use of the Cohen's d test.

Id.

B

This case begins where *Stupp* ended—Was it unreasonable for Commerce to use Cohen's d test as part of its differential pricing analysis when the test is applied to data sets that do not satisfy the statistical assumptions (normal distribution, equal variability, and equally and sufficiently numerous data)? It was.

Commerce makes two arguments to justify its use of Cohen's d test: (1) "[T]he assumptions discussed in *Stupp* are unnecessary when Commerce uses a *full population* of prices in each of the test and comparison groups in its Cohen's d test, rather than *sampling* the sales prices," U.S. Br. 53 (emphasis added); and (2) "Commerce does not rely on the Cohen's d test to calculate the respondent's weighted-average dumping margin, but only to determine whether prices differ significantly as one component of its analysis to determine whether it is appropriate to use an alternative comparison method instead of the A-to-A method in calculating the dumping margin," U.S. Br. 55. This second argument apparently suggests that it does not matter if the calculated Cohen's d coefficient is accurate or not because it is not actually used in the calculation of the dumping margin. *See id.*; WTTC Br. 51–52.

We begin by rejecting Commerce's second argument. As explained above, Cohen's d test is used as the first step in the differential pricing analysis. While we recognize that Cohen's d coefficient is not a direct input into calculating the dumping margin, its value is a basis for determining which methodology to use to calculate the dumping margin. If the output of Cohen's d test is incorrect, then the flawed output becomes a flawed input to the ratio test. That flawed input to the ratio test leads to a flawed output,

and that flawed output will be used to determine whether a pattern of significant price differences exists. And if the ratio test determines there is a pattern when there is not, Commerce has then selected the wrong weight-averaged methodology as required under § 1677f-1(d)(1)(A)(i). We decline to conclude that a flawed output in step one of a calculation can be transformed into meaningful data by further manipulating it.

As to Commerce's first argument regarding samples versus populations, Commerce's argument strains credibility. First, Professor Cohen, who first established Cohen's *d* test, expressly discussed the effect size in terms of a population:

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.

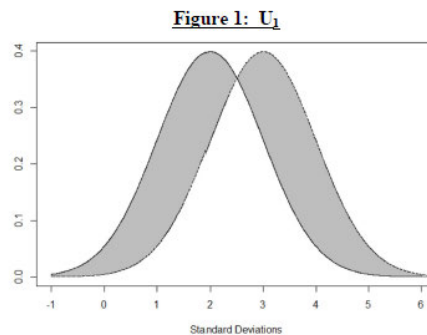
JACOB COHEN, STATISTICAL POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES 21 (2d ed. 1988) (*italics added; bold in original*) (“*Statistical Power Analysis*”).

LARGE EFFECT SIZE: **d** = .8. When our *two populations* are so separated as to make **d** = .8, almost half (**U**₁ = 47.4%) of their areas are not overlapped. **U**₂ = 65.5%, i.e., the highest 65.5% of the *B population* exceeds the lowest 65.5% of the *A population*. As a third measure, the mean or upper half of the *B population* exceeds the lower 78.8% (= **U**₃) of the *A population*.

Id. at 26 (*italics added; bold in original*).

Second, whether the data comes from a sample or a population is irrelevant to producing an accurate Cohen's *d* coefficient. “Cohen's *d* coefficient is a measure of ‘effect size’ that gauges the extent of the difference between the

means of two groups.” U.S. Br. 50; *see also Stupp*, 5 F.4th at 1346–47; Canada Amicus Br. 11. “Effect size measures the ‘effect’ that being in one group rather than the other has on the observed value and expresses that measurement in units of standard deviation.” Canada Amicus Br. 11. For example, Cohen’s d coefficient could be used to measure the effect (i.e., the “effect size”) on test scores (i.e., observed or measured value) from being placed in two different classes. *See id.* For the effect size to have meaning, it is necessary to “maintain the assumption that the populations being compared are normal and with equal variability, and conceive them further as equally numerous.” *Statistical Power Analysis*, at 21. Then, “it is possible to define measures of nonoverlap (U) associated with d which are intuitively compelling and meaningful.” *Id.* (emphasis in original). Essentially, each d value corresponds to a set of U values (U_1 , U_2 , and U_3), which describe how much of the two compared groups overlap with each other. For example, U_1 describes the percentage of all observations in the two groups that do not overlap (the gray areas in the image below).



Canada Amicus Br. 13.

Continuing with this example, when Cohen’s $d = 0.8$, that means $U_1 = 47.4\%$ (i.e., almost half the areas of the groups do not overlap). But the relationship between d values and U values is based on the assumption that the data

sets being compared have normal distributions. *Statistical Power Analysis*, at 23 (“[The U values] are simply related to d and each other through the cumulative normal distribution.” (emphasis in original)). This is so because normal distributions are defined by their mean and standard deviations. Definitionally, in a normal distribution, approximately 68% of data falls within one standard deviation and approximately 95% of data falls within two standard deviations. Thus, whether in a population or a sample, where two groups of data are not normally distributed, the relationships between standard deviation, normal distributions, d values, and U values fall apart. See Canada Amicus Br. 17–19 (demonstrating the flawed outcome in calculating a d coefficient for groups of data that are not normally distributed). In other words, when the data sets do not satisfy the normal assumptions, the Cohen’s d value no longer represents the effect size of the two groups being compared, i.e., how significant is the difference between being in one group and being in the other, specifically regarding the groups’ means.

Commerce offers two main arguments to suggest that the academic literature favors its position. We are not persuaded. First, Commerce argues that because it relies on the entire universe of data (and not a sample), “Commerce uses the Cohen’s d test to measure the *practical* significance of differences in real-world pricing, rather than the *statistical* significance (which arises when one seeks to determine the likelihood that a result observed based on estimation through sampling is a result of chance, or represents the actual parameters of the full populations).” U.S. Br. 54 (emphasis in original) (citing J.A. 4835–37). But the portion of the appendix that Commerce relies on for support is based on a totally different premise—i.e., that a different test, called the “t-test,” tests for statistical significance, while Cohen’s d test, which measures an effect size, measures for a practical significance. See J.A. 4835–37. That there are two different tests for

analyzing whether the means of two sets of data are different says nothing about why the assumptions for Cohen's d test need not be satisfied when relying on a population instead of a sample.

Next, Commerce argues that “unlike with a sample of data for which the estimated parameters will change with each sample selected from a population, each time these parameters are calculated as part of Commerce's Cohen's d test the exact same results will [be] obtain[ed] because the calculated parameters are the actual values of the parameters of the entire population and not estimated values of the parameters based on a sample.” U.S. Br. 54–55. This argument misses the point. Even if you calculate the same mean, standard deviation, and d value with population data (whereas there might be minor variations based on estimated values of sampled data), the d value while consistent could still be unsound as a gauge of group difference. Repeatedly calculating an incorrect d value does not transform it into a “compelling and meaningful” d value. *Statistical Power Analysis*, at 21.

For the reasons above, and those explained in *Stupp*, we conclude that it was unreasonable to rely on Cohen's d test to determine whether prices differ significantly when the underlying data is not normally distributed, equally variable, and equally and sufficiently numerous. Because there is no dispute that Marmen's data does not satisfy these assumptions, Cohen's d test cannot be used here to determine “a pattern of export prices (or constructed export prices) for comparable merchandise that differ significantly among purchasers, regions, or periods of time.” § 1677f-1(d)(1)(B). We therefore vacate Commerce's calculated dumping margin based on the unreasonable use of Cohen's d test to justify the A-to-T methodology. On remand, Commerce may re-perform a differential pricing analysis, and that analysis may not rely on Cohen's d test for data sets like those here. This conclusion, of course, does not preclude Commerce from fashioning and

justifying a statistical analysis that uses some of the ideas underlying Cohen's analysis of group differences as long as the resulting analysis is itself justified as sound for gauging differences in the data sets at issue.

CONCLUSION

We have considered the parties' remaining arguments and find them unpersuasive. For the reasons above, we vacate and remand for Commerce to recalculate Marmen's dumping margin consistent with this opinion.

VACATED AND REMANDED

COSTS

No costs.