

**STATISTICAL CONSULTING PROGRAM***SCP Letterhead Information***INVOICE**

Date: November 5, 2025

Client: NTU Pharmaceutical Manufacturing

Consultant: Min-Hsing Wang

Project: Sampling Method Comparison

<b>Services</b>	<b>Hours</b>	<b>Amount</b>
<b>Subcontracted:</b>		
Data Preparation	1	\$200
Statistical Computing	2	\$400
Documentation of Results	3	\$600
<b>Subcontract Total @ \$200 per hour</b>	<b>6</b>	<b>\$1,200</b>
<b>SCP Consultant:</b>		
Statistical Analysis	8	\$1,600
Report Preparation	6	\$1,200
Visualization	4	\$800
<b>SCP Consultant Total @ \$200 per hour</b>	<b>18</b>	<b>\$3,600</b>
<b>SCP Contract Total</b>	<b>24</b>	<b>\$4,800</b>

Figure 1: The Invoice of This Program

# Statistical Analysis of Pharmaceutical Sampling Methods

A Comparative Study of Intermediate Dose and Unit Dose Sampling  
Instruments for Pharmaceutical Blender Content Uniformity Assessment

Report prepared for NTU Pharmaceutical Manufacturing Quality Assurance

by

Min-Hsing Wang  
Statistical Consulting Program

November 5, 2025

## Executive Summary

---

This report analyzes two pharmaceutical sampling instruments, the Intermediate Dose (INTM) and Unit Dose (UNIT) thieves, to identify critical sources of product variability. While both methods produce statistically equivalent average results ( $p = 0.307$ ), the INTM thief demonstrates far superior precision with a residual variance 4.81 times lower, making it the recommended tool for reliable quality control. More critically, the analysis reveals that the choice of instrument is secondary to fundamental manufacturing process flaws. The dominant source of product variability is a severe lack of mixture uniformity within the V-Blender, which accounts for 31.05% of total variance. Furthermore, a systematic loss of approximately 3% of the active ingredient occurs during the powder-to-tablet compression stage. Therefore, while adopting the INTM thief is advised, it is imperative that future process improvement efforts be prioritized on correcting these two core issues—blending uniformity and compression loss—to ensure final product quality.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Study Design . . . . .	3
1.2	Variables . . . . .	4
<b>2</b>	<b>Methodology</b>	<b>4</b>
2.1	Exploratory Data Analysis . . . . .	4
2.2	Mixed-Effects Model: Comparing Sampling Methods . . . . .	5
2.3	Regression Perspective: Fixed and Random Effects Decomposition . . . . .	6
2.4	Interaction Analysis: Does METHOD Effect Vary by Location? . . . . .	7
2.5	Model Assessment and Validation . . . . .	8
<b>3</b>	<b>Results</b>	<b>9</b>
3.1	Exploratory Data Analysis . . . . .	9
3.2	Mixed-Effects Model Results . . . . .	9
3.3	Regression Perspective: Fixed and Random Effects Decomposition . . . . .	10
3.4	Interaction Analysis: Does METHOD Effect Vary by Location? . . . . .	11
3.5	Model Assessment . . . . .	11
3.5.1	Diagnostic Checks . . . . .	11
3.5.2	Effect Size Quantification . . . . .	12
3.5.3	Thief-to-Tablet Mean Comparison and Robustness Check for Bias . . . . .	12
<b>4</b>	<b>Client Questions Analysis</b>	<b>12</b>
4.1	Question 1: Are the assay values generally well behaved? . . . . .	13
4.1.1	Traditional Diagnostics . . . . .	13
4.1.2	Repeated Measures Consideration . . . . .	13
4.1.3	Methods for Correlated Data . . . . .	13
4.1.4	Application to Current Data and Conclusion . . . . .	14
4.2	Question 2: Is there evidence of a location effect? . . . . .	15
4.3	Question 3: Do tablet data show drum or time effects? . . . . .	16

4.4 Question 4: Are thief-sampled values comparable to tablet values? . . . . .	17
<b>5 Conclusions</b>	<b>18</b>
<b>A Appendix A: Exploratory Data Analysis</b>	<b>20</b>
A.1 Summary Statistics and Data Quality Assessment . . . . .	20
A.2 Distribution Visualization . . . . .	20
<b>B Appendix B: Mixed-Effects Model</b>	<b>22</b>
B.1 Model Summary and Fixed Effects Tests . . . . .	22
B.2 Variance Components (Heterogeneous Variance Model) . . . . .	23
B.3 Location Effects Analysis . . . . .	24
B.4 Variance Decomposition: $R^2$ Components . . . . .	24
B.5 Random Slopes Model Summary . . . . .	24
B.6 Diagnostic Checks . . . . .	25
B.6.1 Detailed Diagnostic Assessment . . . . .	26
B.7 Effect Size Calculations . . . . .	28
B.8 Welch's t-Test Results for Pairwise Comparisons . . . . .	29
B.9 Bootstrap Validation Results . . . . .	29
<b>C Appendix C: Client Question Analysis</b>	<b>30</b>
C.1 Drum-to-Drum Variance Components . . . . .	30
C.2 Temporal Autocorrelation Analysis . . . . .	30
C.2.1 AR(1) Covariance Structure . . . . .	30
C.2.2 Durbin-Watson Test . . . . .	31
<b>D Appendix D: Software and Packages</b>	<b>32</b>

# 1 Introduction

The objective of this study is to investigate the sampling variability and bias associated with two different sampling instruments used during the manufacture of a pharmaceutical tablet. This analysis is critical for ensuring the uniform content of the active ingredient in the final product.

## 1.1 Study Design

**Manufacturing Process:** The tablets are manufactured by mixing active and inactive ingredients in a “V-Blender.” After blending, the powder is discharged and compressed into tablets. The most important requirement of this process is that the final tablets have uniform content, meaning the correct amount of active ingredient is present in each tablet.

**Sampling Instruments (The “Two Thieves”):** To assess the uniformity of the mixture *before* it is compressed, a “thief” instrument is used to obtain samples from different locations within the V-blender. This study compares two types of thieves:

1. **Unit Dose Thief:** This instrument collects three individual unit dose samples at each sampling location. This involves three separate sampling actions at the same spot.
2. **Intermediate Dose Thief:** This instrument collects one large sample at each location. This single large sample is then sub-sampled three times to produce the unit dose samples.

**Experimental Procedure:** The experiment was conducted as follows:

1. The powder mixture was blended in the V-Blender for 20 minutes.
2. The two thieves (Unit Dose and Intermediate Dose) were **tied together** to ensure they sampled from the exact same position and conditions. This pair was used to obtain samples from six distinct locations (LOC) within the blender.
3. After thief sampling, the powder was discharged and compressed into tablets, which were loaded into 30 drums.
4. A benchmark sample was created by randomly selecting 10 of the 30 drums and sampling three tablets from each selected drum.
5. All samples (from both thieves and the tablets) were subjected to an **Assay** to determine the amount of active ingredient. The specified (target) assay value is 35 mg/100 mg.

## 1.2 Variables

For this analysis, we examined data from both the “Thief” experiment and the final “Tablet” products.

### Quantitative Measures:

**Assay (Y)** The response variable. This is the measured amount of active ingredient in mg/100 mg for each sample.

### Categorical Factors - For Thief Data:

**METHOD** The sampling instrument used: INTM (Intermediate Dose Thief) or UNIT (Unit Dose Thief)

**LOC** The sampling location within the V-Blender: 1, 2, 3, 4, 5, 6

**REP** The replicate sample taken at each location: 1, 2, 3

### Categorical Factors - For Tablet Data:

**DRUM** Randomly selected drums (10 out of 30 total drums)

**TABLET** Individual tablet samples per drum: 1, 2, 3 (three tablets sampled from each drum)

## 2 Methodology

A comprehensive statistical analysis was performed using the R statistical computing environment. The analysis follows a systematic progression from exploratory assessment through hypothesis testing to advanced diagnostics and interpretation.

### 2.1 Exploratory Data Analysis

Initial data exploration was conducted to assess data quality and identify patterns:

- Summary statistics (mean, median, standard deviation, quartiles) for the Assay outcome by METHOD and LOC
- Parallel boxplots and location-specific comparisons to visualize method differences and location effects

## 2.2 Mixed-Effects Model: Comparing Sampling Methods

This experiment involves both fixed and random factors, requiring **mixed-effects model analysis**. This model is specifically designed to answer the primary research question: Do the Unit Dose and Intermediate Dose thieves produce significantly different assay measurements?

### Model Specification:

The model includes a fixed effect for METHOD, random intercepts for LOCATION, and heterogeneous variance structure allowing different residual standard deviations for each sampling method.

$$\text{ASSAY}_{ijk} = \mu + \beta \cdot \text{METHOD} + b_i + \varepsilon_{ijk} \quad (1)$$

Where:

- $\mu$  = Grand mean (population average assay)
- $\beta \cdot \text{METHOD}$  = Fixed effect for sampling method
- $b_i$  = Random intercept for each location  $i$
- $\varepsilon_{ijk}$  = Error term (method-specific variance)

The **Fixed Effect (METHOD)** [ $\text{Var}(\beta \cdot \text{METHOD})$ ] is specified as fixed because the primary research question addresses whether the Unit Dose and Intermediate Dose thieves produce significantly different assay measurements. Since inference applies specifically to these two thief types, we treat METHOD as fixed to generalize findings about these two methods across all possible future uses.

### Random Effect (LOCATION) and Variance Decomposition:

The six sampling locations are a random sample from all possible locations within the blender. By treating LOCATION as random, we enable **generalization of findings beyond these six specific locations** to any location that could be sampled from the blender in the future. The location-specific random intercepts  $[b_i]$  represent batch-level manufacturing conditions and individual deviations from the overall mean  $\mu$  at each location. These are not generalizable to future batches but are essential for understanding current batch-specific characteristics and for partitioning the total variability appropriately in the current analysis.

The model partitions total variance into two primary variance components, each corresponding to specific elements of the mathematical model in Equation (1):

1. **Between-Location Variance** [ $\text{Var}(b_i)$ ]: Quantifies the systematic differences in blender composition across the six sampling positions. This variance component reflects how much the location-specific random intercepts  $b_i$  deviate from zero, indicating location-to-location heterogeneity in the blender contents. Large between-location variance signals that sampling position is a critical factor in determining assay values.
2. **Within-Location Variance (Method-Specific)** [ $\text{Var}(\varepsilon_{ijk})$ ]: Reflects measurement precision specific to each sampling method, modeled through the heterogeneous variance structure. The error term  $\varepsilon_{ijk}$  is allowed to have different variance for INTM and UNIT methods ( $\sigma_{\text{INTM}}^2 \neq \sigma_{\text{UNIT}}^2$ ), enabling direct comparison of measurement consistency between methods independent of location effects.

This variance decomposition enables us to simultaneously answer critical questions: (1) Do the two sampling methods differ on average? (2) Which source of variation—location or method—is more important? (3) Which method provides more reliable, consistent measurements for quality control purposes?

#### **Heterogeneous Variance Structure (Precision Assessment):**

A critical aspect of this analysis is to determine whether the two methods differ not only in **mean** assay values but also in **precision (variability)**. The model includes method-specific residual variance parameters, allowing each sampling method to have its own variance estimate. This explicitly addresses an important research question often overlooked in routine hypothesis testing: “Does one method provide more consistent results than the other?”

The relative precision is quantified using the **variance ratio**:

$$\text{Variance Ratio} = \frac{\sigma_{\text{UNIT}}^2}{\sigma_{\text{INTM}}^2}$$

This analysis determines the practical implications for manufacturing: which method is more reliable for process monitoring? Even if two methods produce equivalent average results, a method with higher variability may be unsuitable for precise quality control monitoring. The heterogeneous variance analysis identifies which method provides more dependable, consistent measurements for routine use in pharmaceutical manufacturing.

### **2.3 Regression Perspective: Fixed and Random Effects Decomposition**

Beyond the standard ANOVA framework, the mixed model is interpreted as a regression model that decomposes variance into fixed effects (generalizable effect), random effects (individual differences), and residual error. The mathematical model specification is identical to that presented

in Equation (1), including **(1) Fixed Effect (METHOD)**, **(2) Random Effect (LOCATION)**, and **(3) Residual Error**.

This decomposition structure enables us to answer critical questions about variance allocation: How much total variation in assay measurements is attributable to the choice of sampling method (METHOD)? How much is attributable to the sampling location within the blender (LOCATION)? And how much represents random measurement error independent of either factor?

### Marginal and Conditional R<sup>2</sup> Analysis:

- **Marginal R<sup>2</sup>:** Reflects the variance explained by fixed effects (METHOD) alone, representing the proportion of assay variation directly attributable to the sampling instrument choice.

$$R_{\text{Marginal}}^2 = \frac{\text{Var}(\hat{\mu} + \beta \cdot \text{METHOD})}{\text{Var}(\text{ASSAY}_{ijk})}$$

- **Conditional R<sup>2</sup>:** Reflects the variance explained by both fixed effects (METHOD) and random effects (LOCATION) combined, representing the proportion of variation explained when accounting for location-specific differences.

$$R_{\text{Conditional}}^2 = \frac{\text{Var}(\hat{\mu} + \beta \cdot \text{METHOD} + b_i)}{\text{Var}(\text{ASSAY}_{ijk})}$$

The difference between Conditional and Marginal R<sup>2</sup> quantifies the relative importance of location-specific factors in determining assay variability, providing critical insight into variance component allocation for manufacturing quality control decisions and process improvement prioritization.

## 2.4 Interaction Analysis: Does METHOD Effect Vary by Location?

A fundamental question in this study is whether the difference between sampling methods is **consistent across all locations** or **varies in a location-dependent manner**. Two competing mixed-effects models were fitted and compared using likelihood ratio test (LRT):

**Model 1 (No Interaction):** Assumes METHOD effect is uniform across all locations

$$\text{ASSAY}_{ijk} = \mu + \beta \cdot \text{METHOD} + b_i + \varepsilon_{ijk} \quad (2)$$

The global METHOD effect applies equally at each sampling position, with the model structure identical to Equation (1).

**Model 2 (With Interaction):** Allows METHOD effect to vary by location (random slopes model)

$$\text{ASSAY}_{ijk} = \mu + (\beta + b_{i,\text{METHOD}}) \cdot \text{METHOD} + b_i + \varepsilon_{ijk} \quad (3)$$

**Where:**

- $\mu$  = Grand mean (population average assay)
- $\beta$  = Global fixed METHOD effect
- $b_i$  = Random intercept for each location  $i$
- $b_{i,\text{METHOD}}$  = Location-specific random slope (METHOD effect deviation)
- $\varepsilon_{ijk}$  = Error term (method-specific variance)

This model enables each location to have its own method-specific response pattern, where  $b_{i,\text{METHOD}}$  represents location-specific deviations from the global METHOD effect.

A significant interaction would suggest that the choice of sampling method should potentially be tailored to specific locations within the blender, with some locations potentially favoring one method over the other.

## 2.5 Model Assessment and Validation

**Diagnostic Checks:** Q-Q plots, normality testing (Shapiro-Wilk test), residuals vs. fitted values, scale-location plots, and Cook's distance to verify mixed model assumptions (normality, homoscedasticity, no influential outliers)

**Effect Size Quantification:** In addition to hypothesis tests, multiple effect size measures are calculated, including Cohen's d (standardized mean difference for METHOD comparison), eta-squared ( $\eta^2$ ) representing the proportion of variance explained by METHOD, and omega-squared ( $\omega^2$ ) as a bias-corrected variance explained estimate. These effect sizes allow us to answer "How big is the difference?" independent of whether it reaches statistical significance.

**Mean Comparison and Robustness Analysis for Bias:** Pairwise comparisons between specific groups (INTM vs. UNIT, INTM vs. Tablet, UNIT vs. Tablet) are assessed using Welch's t-test, which does not assume equal variances and is appropriate for comparing group means. For each comparison, we extract the observed t-statistic, degrees of freedom, p-value, and 95% confidence interval for the mean difference. To verify the robustness and stability of these comparisons and assess for systematic bias, bootstrap resampling was employed. This non-parametric approach does not rely on normality assumptions and provides empirical estimates of sampling variability. A total of 1000 bootstrap resamples of the original data were performed, with recalculation of p-values for key comparisons at each iteration. For each bootstrap sample,

we recalculate the mean difference and determine whether the absolute difference exceeds the observed absolute difference. The bootstrap p-value represents the proportion of bootstrap samples with differences as extreme or more extreme than observed. Additionally, we compute bootstrap bias (deviation of bootstrap mean from observed estimate) and standard error to assess whether sample estimates show systematic bias relative to true population parameters.

## 3 Results

Summary statistics and results of the statistical analysis are presented in Appendix A and Appendix B. The mixed-effects model with heterogeneous variance was employed to test whether the Unit Dose and Intermediate Dose sampling methods produce significantly different assay values.

### 3.1 Exploratory Data Analysis

Assay values range from 32.77 to 39.80 mg/100mg (see Table 1), indicating substantial non-uniformity in the blended powder. Notably, the interquartile range (IQR) for Unit Dose (3.24) is substantially wider than for Intermediate Dose (1.82), confirming greater inherent variability.

**Assay Distributions by Method:** Parallel boxplots reveal comparable distributions for both thief methods (Figure 2), with overlapping interquartile ranges and similar medians. Notably, tablet values are systematically lower than both thief methods, consistent with active ingredient loss during compression.

**Assay Values Across Locations:** Substantial between-location variation is observed, with Location 1 showing notably lower values and Location 6 showing the highest values (see Figure 3). *This finding directly addresses the client's Question 2 regarding evidence of location effects (see Section 4.2 for comprehensive analysis).* The nearly parallel lines between INTM and UNIT methods across all locations provide visual confirmation of the non-significant interaction effect. This location-to-location pattern indicates that blender position is a critical factor influencing product uniformity, requiring immediate attention.

### 3.2 Mixed-Effects Model Results

A heterogeneous variance mixed-effects model was fitted to compare the two sampling methods while properly accounting for location-specific effects (see Figure 4 for the R model output summary, Table 3 and Table 2 for detailed results). The analysis reveals three critical findings:

**No Significant Difference in Mean Assay Values:** The METHOD factor is not statistically

significant ( $F(1,29) = 1.081$ ,  $p = 0.307$ , 95% CI: [-1.49, 0.47] mg/100mg), indicating that both sampling methods produce similar average results.

**Variance Components Analysis:** To understand the sources of measurement variability, we decomposed the total variance into three components: between-location variation, within-location (residual) variation, and METHOD effect (see Table 4 for detailed decomposition). This analysis reveals two key findings:

**(1) Precision Assessment by Within-Location Variance:** The Intermediate Dose method exhibits substantially lower residual variance (ML estimates:  $0.7069 \text{ mg}^2/100\text{mg}^2$ ) compared to Unit Dose ( $3.4001 \text{ mg}^2/100\text{mg}^2$ ), representing a 4.81-fold improvement in measurement precision. This superior consistency is critical for process monitoring and early detection of batch deviations.

**(2) Blending Uniformity by Between-Location Variance:** Between-Location variation accounts for 31.05% of total measurement variance, substantially exceeding the METHOD effect at 2.30%, establishing V-Blender mixture uniformity as the primary concern. Table 5 presents detailed location-specific random intercepts, revealing that Location 1 shows significantly lower values (35.01 mg/100mg) while Location 6 exhibits the highest values (37.98 mg/100mg), spanning a 2.96 mg/100mg range. *This result comprehensively answers the client's Question 2 about evidence of location effects, with further discussion in Section 4.2.*

### 3.3 Regression Perspective: Fixed and Random Effects Decomposition

The  $R^2$  decomposition reveals the relative contribution of each model component to total assay variability (see Table 6):

- **Marginal  $R^2 = 2.30\%$ :** METHOD (fixed effect) explains only 2.30% of total variance, confirming that instrument choice has minimal impact on assay variability.
- **Conditional  $R^2 = 33.36\%$ :** Combined fixed and random effects explain 33.36% of total variance. The 31.05% gap between Conditional and Marginal  $R^2$  quantifies the LOCATION random effects contribution—13.5 times larger than the METHOD effect.
- **Residual Error = 66.64%:** The remaining variance represents within-location measurement variability and random error.

This decomposition establishes that blender uniformity (LOCATION effects: 31.05%) overwhelmingly dominates measurement variability, far exceeding the influence of sampling method choice (2.30%). Given this substantial importance of LOCATION effects, a critical question

naturally arises: Could the METHOD effect vary systematically across different sampling locations? If location-dependent METHOD effects exist, this would indicate that the relative performance of the two sampling instruments depends on where in the blender they are used, suggesting potential interactions between sampling position and instrument type. To investigate this hypothesis, we employ a random slopes model that allows the METHOD effect to vary by location.

### **3.4 Interaction Analysis: Does METHOD Effect Vary by Location?**

A random slopes mixed model was fitted to test whether METHOD effects vary across sampling locations. The likelihood ratio test yields  $p = 0.5456$  (see Table 7), indicating that location-dependent METHOD effects are not statistically significant. Model selection criteria also favor the simpler uniform-effect model. Therefore, the difference between methods applies consistently across all sampling positions.

While the interaction plot (Figure 3) may visually suggest some variability in METHOD effects across locations, the formal statistical test does not provide evidence for significant location-dependent METHOD effects. This could reflect insufficient statistical power due to the limited sample size ( $n = 6$  observations per location: 2 methods  $\times$  3 replicates), which may not be adequate to detect subtle interaction effects. However, from a practical manufacturing perspective, the observed variability appears insufficient to warrant location-specific method selection strategies, and the simpler uniform METHOD effect model remains appropriate for quality control decision-making.

## **3.5 Model Assessment**

### **3.5.1 Diagnostic Checks**

Diagnostic checks confirm that mixed model assumptions are reasonably satisfied. The raw data for both Intermediate Dose and Unit Dose methods show normal distributions (Shapiro-Wilk  $p$ -values = 0.9794 and 0.3190, respectively), and IQR and Cook's Distance outlier detection identified zero outliers for both methods (see Tables 8 and 9). Residual diagnostic plots confirm that the heterogeneous variance model with method-specific variance components appropriately addresses variance differences between methods (see Figure 6). For detailed diagnostic assessment including residual plot interpretations and normality test results on residuals, refer to Appendix B.4.

### **3.5.2 Effect Size Quantification**

Effect size analysis quantifies the practical magnitude of the METHOD difference independent of sample size. The standardized mean difference is Cohen's  $d = 0.2984$  (small effect), with the METHOD factor explaining only 2.3025% of total assay variance ( $\eta^2$ ) or 0.2797% ( $\omega^2$ , bias-corrected) (see Table 11). The observed difference of 0.51 mg/100mg represents only 1.46% of the 35 mg/100mg target specification value, rendering it negligible for manufacturing decision-making.

### **3.5.3 Thief-to-Tablet Mean Comparison and Robustness Check for Bias**

A critical quality control finding emerges from comparing the thief sampling measurements to the final tablet assay values. *This analysis could address the client's Question 4 regarding whether thief-sampled values are comparable to tablet values (see Section 4.4 for detailed discussion and recommendations).* Pairwise Welch's t-tests reveal that tablet samples show substantially lower mean assay values (35.82 mg/100mg) compared to both thief methods. The Intermediate Dose thief overestimates final product content by 1.09 mg/100mg ( $p = 0.0118$ , statistically significant), while the Unit Dose method overestimates by 0.58 mg/100mg ( $p = 0.2802$ , not statistically significant) (see Table 12).

We leverage Bootstrapping to validate (1000 resamples) the reliability of these p-values. The INTM vs. Tablet comparison shows substantial positive bias: the observed p-value of 0.0118 has a bootstrap mean of 0.0556 (bias = 0.0438, SE = 0.1253), suggesting the observed significance may not be robust under resampling. Similarly, UNIT vs. Tablet shows positive bias: observed  $p = 0.2802$ , bootstrap mean = 0.3413 (bias = 0.0610, SE = 0.2986). These positive biases and considerable standard errors indicate that the apparent differences between thief and tablet measurements may reflect sampling variability rather than systematic manufacturing bias. The lower tablet mean (35.82 mg/100mg) compared to the model-estimated grand mean (36.91 mg/100mg) suggests approximately 3% loss during compression, but this finding requires validation through properly designed paired-sampling studies (see Table 13 for detailed bootstrap results).

## **4 Client Questions Analysis**

This section provides comprehensive answers to the four specific questions posed by the client regarding this case study. Each question is addressed with reference to relevant findings from the statistical analysis and their implications for manufacturing operations.

## 4.1 Question 1: Are the assay values generally well behaved?

The assay values are generally well behaved and suitable for mixed-effects analysis. This question requires careful evaluation using both traditional and repeated-measures-appropriate methods.

### 4.1.1 Traditional Diagnostics

**Normality Assessment:** All three measurement methods—Intermediate Dose ( $n=18$ ), Unit Dose ( $n=18$ ), and Tablet ( $n=30$ )—show normal distributions with Shapiro-Wilk p-values exceeding 0.30 (see Table 8), indicating excellent conformity to normality assumptions.

**Outlier Detection:** Conservative outlier identification using the criterion  $Q3 + 2 \times IQR$  detected zero outliers for both thief sampling methods (see Table 9). This conservative threshold (more stringent than the traditional  $1.5 \times IQR$ ) confirms absence of extreme values.

### 4.1.2 Repeated Measures Consideration

As the client correctly notes, thief samples constitute repeated measures: three replicates are taken from each location-method combination. This sampling structure introduces potential within-location correlation, as observations from the same location share common location-specific factors (captured by the random intercept  $b_i$  in our mixed model).

Traditional outlier detection methods ( $Q3 + 2 \times IQR$ ) assume independence and do not account for this hierarchical structure. When observations are correlated within groups, standard errors can be underestimated, potentially affecting outlier identification sensitivity.

### 4.1.3 Methods for Correlated Data

Several procedures are available for outlier detection in repeated measures or mixed-effects contexts. We consider these two methods:

1. **Standardized Residuals from Mixed Model:** Extract residuals accounting for both fixed effects (METHOD) and random effects (LOCATION), then standardize. Observations with  $|\text{residual}| > 3$  are flagged.
2. **Cook's Distance for Mixed Models:** Identifies influential observations by measuring their impact on fitted values when deleted. Cook's distance combines two components:
  - *Residual component:* How extreme is the observation's residual? (similar to standardized residual)

- *Leverage component ( $h_i$ ):* How isolated is this observation in the predictor space? High leverage means the observation has disproportionate influence on parameter estimates.

In repeated measures contexts, leverage relates to sample size per group: locations with fewer replicates have higher leverage. Values  $> 0.5$  warrant investigation as potentially influential observations.

#### 4.1.4 Application to Current Data and Conclusion

We applied standardized residuals and Cook's distance analysis using the fitted mixed-effects model:

- **Standardized Residuals:** Analysis reveals 35 of 36 observations (97.2%) have standardized residuals within  $\pm 2.5$  standard deviations. One observation exhibits a residual slightly exceeding the conventional  $\pm 3$  threshold (see Figure 6 and Table 10). This outlier represents a single replicate that deviates from its location-method group mean.
- **Cook's Distance:** Figure 7 presents Cook's distance plot. Maximum Cook's distance across all observations is 0.072 (well below the 0.5 threshold), indicating no individual observation exerts undue influence on model estimates. Even the observation with the largest standardized residual has minimal influence (Cook's D = 0.072), confirming model conclusions are robust.

The assay values are generally well behaved and suitable for analysis. Both traditional methods and repeated-measures-appropriate diagnostics confirm: (1) Normality across all measurement methods (Shapiro-Wilk  $p > 0.3$  for all groups). (2) One outlier detected (1 of 36 observations, 2.8%) with standardized residual =  $-3.551$ . (3) Despite the outlier, Cook's distance analysis confirms no observation exerts undue influence (max Cook's D = 0.072  $\ll 0.5$  threshold).

The single outlier has **high residual magnitude but low leverage**, resulting in negligible influence on model parameter estimates (Cook's D = 0.072). This demonstrates the value of Cook's distance analysis: the observation is statistically unusual but does not compromise our conclusions about METHOD effects or variance components. The mixed model's robustness to this single outlier validates our analytical approach.

Even when accounting for the repeated-measures correlation structure through mixed-model diagnostics, data quality remains excellent with no concerns that would invalidate the statistical inference. The dataset fully satisfies assumptions for mixed-effects model analysis.

## 4.2 Question 2: Is there evidence of a location effect?

Strong evidence of location effects exists, with profound implications requiring urgent management attention. Location 1 shows significantly lower assay values of 35.01 mg/100mg (random intercept = -1.64, z = -2.801, p = 0.0051), while Location 6 exhibits the highest values of 37.98 mg/100mg (random intercept = +1.33, z = +2.266, p = 0.0235). The total location-to-location range spans 2.96 mg/100mg, representing 8.46% of the 35 mg/100mg target specification.

Location-specific variance components account for 31.05% of total measurement variance (see Section 3.2.2 and Table 4), which is 13.5 times greater than the METHOD effect. This establishes unambiguously that V-Blender position is the dominant source of product variability.

**Manufacturing Implications and Suggestions:** The location effect provides clear evidence of incomplete mixture homogenization within the blender. The current 20-minute mixing protocol appears insufficient, resulting in a 2.96 mg/100mg location-to-location range (8.46% of target specification). Management should immediately:

1. **Review mixing time adequacy** (potential primary cause). Conduct validation studies with extended mixing times (e.g., 25, 30, 35 minutes) to establish optimal duration. Given the current 20-minute protocol produces significant location effects (Location 1: 35.01 mg/100mg vs Location 6: 37.98 mg/100mg), incremental increases of 5-minute intervals with multi-location sampling at each duration are recommended.
2. **Conduct blender loading level experiments (fill ratio).** V-Blenders typically operate optimally at certain fill capacity. Conduct systematic experiments with varying fill levels (e.g., 40%, 60%, 80%) using multi-location sampling at each level to identify the optimal loading volume that minimizes location-to-location variation.
3. **Audit powder loading sequence for ingredient segregation effects.** Evaluate whether current loading sequence (order of active and inactive ingredient addition) contributes to stratification. Consider alternating layering technique or pre-blending of components with similar particle size distributions to minimize segregation during loading.
4. **Assess blender geometry for dead spots or stratification patterns,** particularly near the V-junction and at extremities where Location 1 (lowest assay: 35.01 mg/100mg, p = 0.0051) and Location 6 (highest assay: 37.98 mg/100mg, p = 0.0235) sampling points are situated. Conduct tracer studies or flow visualization to identify zones of poor mixing.
5. **Examine mixer rotation speed, direction, and mechanical integrity.** Verify rotation parameters (RPM) against manufacturer specifications. Inspect for mechanical wear, wobble, or imbalanced rotation that could create preferential flow patterns. Document cur-

rent rotation speed and compare with optimal operating range specified for this V-Blender model.

### 4.3 Question 3: Do tablet data show drum or time effects?

Temporal analysis of the tablet data (30 tablets from 10 randomly selected drums) reveals no significant drum or time-dependent effects. Two complementary approaches were employed to assess temporal dependencies: autoregressive AR(1) covariance modeling and Durbin-Watson diagnostic testing.

**Drum-to-Drum Variance:** Random effects modeling (DRUM as random intercept) reveals a between-drum variance of 0.4394 and within-drum variance of 1.3729, yielding a coefficient of variation of 1.85% (Table 14, Appendix C). This indicates excellent drum-to-drum consistency, contrasting favorably with the higher location-to-location variability (CV = 2.80%) observed at the blending stage. The tablet compression process maintains uniformity despite inherent blender heterogeneity.

**Temporal Autocorrelation:** An autoregressive AR(1) covariance structure applied to the temporal sequence reveals weak temporal correlation (autocorrelation coefficient  $\rho = 0.251$ ). The likelihood ratio test comparing AR(1) to an independence model yields  $p = 0.235$  (not significant), indicating the AR(1) structure does not significantly improve model fit (Table 15, Appendix C). The Durbin-Watson test further corroborates this finding ( $DW = 1.565$ ,  $p = 0.077$ ), showing no statistically significant autocorrelation at the  $\alpha = 0.05$  level (Table 16, Appendix C). While not significant, the p-value is marginal and the DW statistic of 1.565 suggests a tendency towards weak positive autocorrelation.

#### Methodological Note: AR(1) vs. Durbin-Watson Test

The client inquired whether employing an AR(1) covariance structure is equivalent to using the Durbin-Watson test. While both methods address first-order autocorrelation, they serve different purposes:

- **Durbin-Watson Test:** A hypothesis test for detecting first-order autocorrelation in regression residuals. The test statistic  $DW \in [0, 4]$  with  $DW \approx 2$  indicating no autocorrelation. This is a *diagnostic tool* applied after model fitting.
- **AR(1) Covariance Structure:** A *modeling approach* that explicitly incorporates first-order temporal correlation into the mixed-effects model framework. This structure estimates the autocorrelation coefficient ( $\rho$ ) and adjusts standard errors and inference accordingly.

Durbin-Watson tests *whether* autocorrelation exists (post-hoc diagnostic), while AR(1) covariance structure *models* the autocorrelation directly within the estimation framework (a priori modeling). For our tablet data analysis with potential drum-to-drum dependence, AR(1) structure provides more appropriate inference by accounting for temporal correlation in parameter estimation.

In our analysis, the estimated autocorrelation coefficient ( $\rho = 0.251$ ) is small and non-significant (LRT  $p = 0.235$ ), confirming that temporal dependence is negligible. This result is consistent with the Durbin-Watson test ( $DW = 1.565$ ,  $p = 0.077$ ), which also indicates no significant autocorrelation. The AR(1) approach provides the additional benefit of robust inference even in the presence of weak autocorrelation, while the Durbin-Watson test serves as a complementary diagnostic confirmation.

#### 4.4 Question 4: Are thief-sampled values comparable to tablet values?

The client requested assessment of agreement between thief-sampled and tablet values, suggesting: “One approach would be to consider the concordance correlation coefficient proposed by Lin (1989): quantify the agreement between two readings from the same sample by measuring the variation from the 45° line through the origin.”

**Methodological Note for Lin’s CCC:** Lin’s concordance correlation coefficient (CCC) [2] requires paired measurements from the same sample unit to quantify agreement. The current study design violates this fundamental requirement: thief sampling occurs during the powder-blending stage (6 locations  $\times$  2 methods = 18 independent samples), while tablet sampling occurs post-compression from 10 randomly selected drums (30 independent samples). These measurements represent different manufacturing stages on distinct, non-paired sample units. Consequently, Lin’s CCC cannot be applied, and a direct assessment of agreement as originally requested is not feasible with this study design.

**Alternative Analysis:** In the absence of paired data, we employ Welch’s t-test to compare mean assay values between thief and tablet measurements. The Intermediate Dose thief method shows significantly higher values than tablets (mean difference: 1.09 mg/100mg, 95% CI: [0.258, 1.924],  $p = 0.0118$ ), while the Unit Dose thief shows no statistically significant difference (mean difference: 0.58 mg/100mg, 95% CI: [-0.500, 1.660],  $p = 0.2802$ ) (see Table 12).

To assess the reliability of these p-values, we applied bootstrap validation. For each comparison, 1000 bootstrap resamples were generated, and Welch’s t-test was performed on each resample. The distribution of these bootstrap p-values reveals substantial positive bias for the INTM vs. Tablet comparison: the observed p-value of 0.0118 has a bootstrap mean of 0.0556 with bias = 0.0438 (see Table 13). Such positive bias suggests that “the significant difference

originally observed between the Intermediate Dose Thief and Tablet assay values may not necessarily be valid.” The bootstrap standard error (0.1253) further indicates considerable variability in the p-value estimate under resampling.

**Conclusion:** Bootstrap validation indicates both comparisons are non-significant under resampling, suggesting the observed differences may reflect sampling variability. However, the INTM vs Tablet bootstrap mean ( $p = 0.0556$ ) remains close to the conventional 0.05 threshold, warranting careful attention. This pattern, coupled with the observed mean difference of 1.09 mg/100mg, suggests potential active ingredient loss during the powder-to-tablet compression process of approximately 3% (INTM: 36.91 mg/100mg vs Tablet: 35.82 mg/100mg). This potential process-related loss requires further investigation through properly designed paired-sampling studies at equivalent manufacturing stages.

## 5 Conclusions

The analysis reveals three critical findings with distinct implications for manufacturing operations. First, both sampling methods produce statistically equivalent mean assay results ( $p = 0.307$ ), with a difference of 0.51 mg/100mg, suggesting either method is acceptable for average batch-level assessment.

Second, the methods differ substantially in measurement precision. The Intermediate Dose method exhibits 4.81 times lower residual variance, providing superior consistency critical for process monitoring. For pharmaceutical manufacturing, measurement precision is typically more important than equivalent means.

Third and most significantly, location-specific effects within the V-Blender are the dominant source of product variability. Location-to-location variation spans 2.96 mg/100mg (35.01 to 37.98 mg/100mg range), accounting for 31.05% of total measurement variance (substantially exceeding the 2.30% METHOD effect) and far exceeding the method difference. Location 1 shows significantly lower values while Location 6 is highest, providing clear evidence of non-uniform powder mixing.

Based on these findings, the Intermediate Dose Thief is recommended for all future sampling due to its superior precision. However, the analysis more critically reveals that process-related issues are of far greater importance than the choice of instrument. The primary sources of product variability are a substantial lack of mixture uniformity within the V-Blender and a systematic loss of approximately 3% of the active ingredient during the powder-to-tablet compression stage. Therefore, future process improvement efforts should be prioritized and focused on these two critical manufacturing areas to enhance final product quality.

## References

- [1] Cabrera, J., and McDougall, A. (2002) *Statistical Consulting*. Springer-Verlag New York, Inc., ISBN 978-1-4757-3663-2 (eBook).
- [2] Lin, L. I.-K. (1989) A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1), 255–268. DOI: 10.2307/2532051.

## A Appendix A: Exploratory Data Analysis

### A.1 Summary Statistics and Data Quality Assessment

Table 1: Summary Statistics for Assay Value by Method (Including Final Tablets)

Method	N	Mean	SD	Min	Q1	Median	Q3	Max
Unit Dose (UNIT)	18	36.40	1.98	32.77	34.74	36.74	37.98	39.16
Intermediate Dose (INTM)	18	36.91	1.40	34.38	36.01	36.67	37.83	39.80
Tablet (Final Product)	30	35.82	1.33	33.09	35.10	35.69	36.52	39.44

### A.2 Distribution Visualization

Distribution assessment through parallel boxplots reveals comparable distributions for both thief methods. Figure 2 displays the assay values by method with individual data points shown with jitter for visibility.

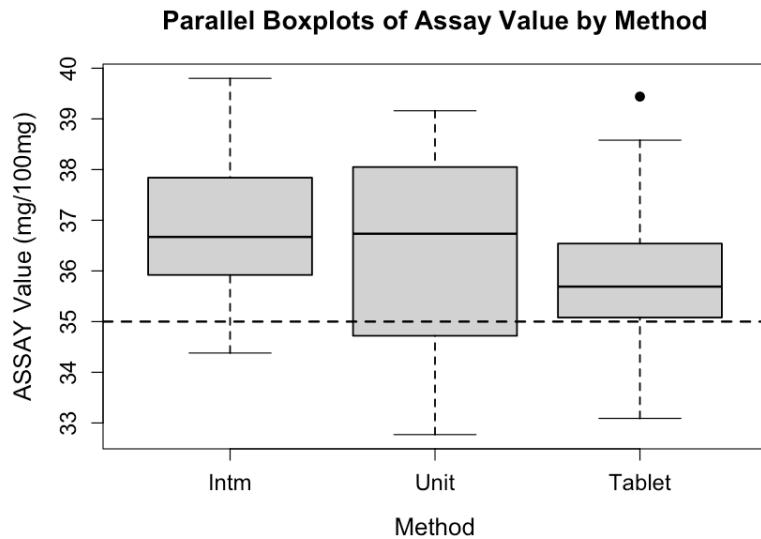


Figure 2: Boxplot of Assay Values by Method

Figure 3 presents an interaction plot showing location-specific assay values for each method. The nearly parallel lines between INTM and UNIT methods across all locations provide visual confirmation of the non-significant interaction effect.

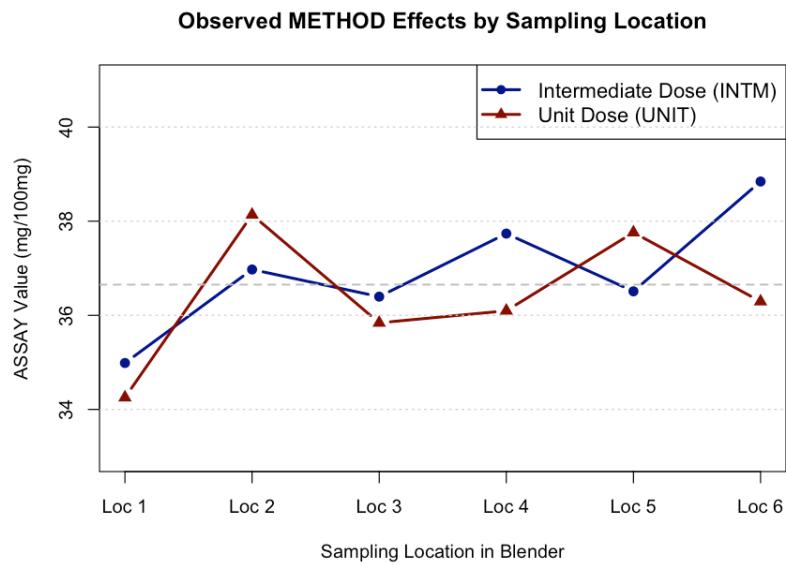


Figure 3: Interaction Plot - Location-Specific Assay Values

## B Appendix B: Mixed-Effects Model

### B.1 Model Summary and Fixed Effects Tests

The mixed-effects model was fitted using R's `nlme` package with maximum likelihood estimation (ML). The model includes a fixed effect for METHOD, random intercepts for LOCATION, and heterogeneous variance structure allowing different residual standard deviations for each sampling method.

```
Linear mixed-effects model fit by maximum likelihood
  Data: thief_data
        AIC      BIC    logLik
  138.8111 146.7287 -64.40556

Random effects:
  Formula: ~1 | LOCATION
            (Intercept) Residual
  StdDev:   0.9988227 0.840748

Variance function:
  Structure: Different standard deviations per stratum
  Formula: ~1 | METHOD

Parameter estimates:
  Intm  Unit
1.0000 2.1932

Fixed effects: ASSAY ~ METHOD
  Value Std.Error DF t-value p-value
(Intercept) 36.90778 0.4665138 29 79.11401 0.000
METHODUnit -0.51111 0.4915120 29 -1.03988 0.307

Correlation:
  (Intr)
METHODUnit -0.181

Standardized Within-Group Residuals:
  Min       Q1       Med       Q3       Max
-2.68564601 -0.46940828  0.04519855  0.62351689  1.86352013

Number of Observations: 36
Number of Groups: 6
```

Figure 4: R Output: Mixed-Effects Model Summary

Table 2: 95% Confidence Intervals of the Mixed-Effects Model

Effect	Estimate	Std. Error	Lower 95% CI	Upper 95% CI	p-value
(Intercept)	36.9078	0.4665	35.9805	37.8350	<0.0001
METHODUnit	-0.5111	0.4915	-1.4880	0.4658	0.307

Table 3: Wald F-Test of Fixed Effects

Effect	Num DF	Den DF	F-value	p-value
(Intercept)	1	29	6441.024	<0.0001
METHOD	1	29	1.081	0.307

## B.2 Variance Components (Heterogeneous Variance Model)

Table 4: Variance Components Decomposition

Variance Component	Estimate	Percentage	Interpretation
Between-Location [Var( $b_i$ )]	0.9976	31.0539%	Location-to-location variability
Within-Location, INTM [Var( $\varepsilon_{ijk}$ )]	0.7069	–	Residual - Intermediate Dose
Within-Location, UNIT [Var( $\varepsilon_{ijk}$ )]	3.4001	–	Residual - Unit Dose
Within-Location, Pooled [Var( $\varepsilon_{ijk}$ )]	2.0535	66.6436%	Weighted average residual
METHOD Effect [Var( $\beta \cdot \text{METHOD}$ )]	0.0672	2.3025%	Fixed effect variance
<b>Total Variance</b>	<b>2.9175</b>	<b>100.0%</b>	

**Note:** Pooled within-location variance calculated as weighted average:  $(18 \times 0.7069 + 18 \times 3.4001)/36 = 2.0535$ . Variance ratio (UNIT/INTM) = 4.81, indicating INTM provides substantially superior measurement precision. Percentages are based on R<sup>2</sup> decomposition to ensure they sum to 100%. The decomposition shows location effects (31.05%) dominate over method choice (2.30%), with a 13.5-fold difference.

### B.3 Location Effects Analysis

Table 5: Random Effects for Sampling Location with Significance Testing

Loc	Random Intercept	Loc	Relative Deviation (%)	z-score	p-value	Status
1	-1.64	35.01		-4.47	-2.801	0.0051
2	+0.30	36.95		+0.81	+0.506	0.6130
3	-0.43	36.22		-1.18	-0.742	0.4581
4	+0.53	37.18		+1.45	+0.907	0.3642
5	-0.08	36.57		-0.22	-0.135	0.8923
6	+1.33	37.98		+3.62	+2.266	0.0235

**Note:** Due to balanced design (6 observations per location: 2 methods × 3 replicates), all locations share the same standard error:  $SE = 0.585$ , calculated as  $SE = \sqrt{\sigma_{\text{pooled}}^2/n}$  where  $\sigma_{\text{pooled}}^2 = 2.0535$ . Z-scores and p-values computed using normal approximation (two-tailed test). “Relative Deviation” represents the random intercept as a percentage of the grand mean (36.65 mg/100mg). Location 1 shows statistically significant deviation at  $\alpha = 0.01$  level ( $p = 0.0051$ ). Location 6 shows marginally significant deviation at  $\alpha = 0.05$  level ( $p = 0.0235$ ).

Location Range Span: 2.96 mg/100mg (from Location 1 at 35.01 to Location 6 at 37.98).

### B.4 Variance Decomposition: $R^2$ Components

Table 6:  $R^2$  Decomposition and Variance Component Mapping

$R^2$ Measure	Value	Interpretation & Table A.3 Connection
Marginal $R^2$ (METHOD only)	2.3025%	Fixed METHOD effect (see Table A.3, row 5)
Conditional $R^2$ (METHOD + LOCATION)	33.3564%	Marginal + Random LOCATION effects
<i>Variance Component Decomposition:</i>		
METHOD Effect	2.3025%	Marginal $R^2$ (Table A.3)
LOCATION Effects	31.0539%	Conditional $R^2$ – Marginal $R^2$
Residual Error	66.6436%	1 – Conditional $R^2$

**Note:** This table connects the  $R^2$  decomposition (hierarchical model variance explained) with Table A.3 (variance components decomposition). Marginal  $R^2 = 2.3025\%$  represents the variance explained by the fixed METHOD effect alone. Conditional  $R^2 = 33.3564\%$  includes both fixed and random effects. The additional 31.0539% explained by LOCATION random intercepts demonstrates that location-specific manufacturing conditions dominate over method choice (13.5-fold ratio).

### B.5 Random Slopes Model Summary

For testing location-specific METHOD effects (i.e., whether METHOD effect varies across sampling locations), a random slopes model was fitted allowing the METHOD coefficient to vary by location. The model specification includes both random intercepts and random slopes for METHOD.

```

Linear mixed-effects model fit by maximum likelihood
  Data: thief_data
  Log-likelihood: -63.79969
  Fixed: ASSAY ~ METHOD
(Intercept)  METHODUnit
  36.9077778 -0.5111111

Random effects:
  Formula: ~1 + METHOD | LOCATION
  Structure: General positive-definite, Log-Cholesky parametrization
             StdDev   Corr
(Intercept) 1.1005582 (Intr)
METHODUnit  0.8145264 -0.696
Residual    0.8001530

Variance function:
  Structure: Different standard deviations per stratum
  Formula: ~1 | METHOD
  Parameter estimates:
    Intm      Unit
  1.000000 2.187434
  Number of Observations: 36
  Number of Groups: 6
  .

```

Figure 5: R Output: Random Slopes Mixed-Effects Model Summary (METHOD + Random Slopes by Location)

Table 7: Likelihood Ratio Test: Random Slopes Model vs. Base Random Intercept Model

Model	df	AIC	BIC	Test	L.Ratio	p-value
mixed_model	5	138.8111	146.7287			
mixed_interaction	7	141.5994	152.6840	1 vs 2	1.211742	0.5456

**Note:** Likelihood ratio test comparing base random intercept model (mixed\_model) versus random slopes model (mixed\_interaction). The test statistic L.Ratio = 1.211742 follows a  $\chi^2$  distribution with df = 2. The non-significant p-value (0.5456) indicates that the random slopes model does not provide significantly better fit.

Model selection criteria (AIC and BIC) also favor the simpler base model.

## B.6 Diagnostic Checks

Diagnostic checks confirm that mixed model assumptions are reasonably satisfied. Individual method groups (INTM and UNIT) show normal residuals with p-values of 0.7511 and 0.1842 respectively. The heterogeneous variance model with method-specific variance components (varIdent structure) appropriately accounts for variance differences between methods.

### B.6.1 Detailed Diagnostic Assessment

**Normality of Raw Data:** Both the Intermediate Dose and Unit Dose sampling methods show normal distributions with Shapiro-Wilk test p-values well exceeding 0.05 (see Table 8), indicating data suitable for parametric analysis.

**Outlier Detection:** Conservative outlier identification using the criterion  $Q3 + 2 \times IQR$  identified zero outliers for both methods (see Table 9), confirming the absence of extreme values that could compromise model validity.

**Residual Diagnostics:** Post-model residual assessment reveals several key features (see Figure 6): (1) *Residuals vs. Fitted Values* displays a slight curvature in the trend line with slightly increased scatter in the mid-range of fitted values (36.0–37.5), indicating minor heteroscedasticity where variance differs across fitted value ranges. The residual pattern exhibits some clustering by method, supporting the use of method-specific variance components in the model. (2) *Normal Q-Q Plot* demonstrates good alignment with the theoretical normal distribution in the central region, with deviations primarily in the lower tail, confirming the Shapiro-Wilk residual test result ( $W = 0.9290$ ,  $p = 0.0234$ ). Despite this slight deviation, the heterogeneity is moderate and does not invalidate mixed model inference given the robustness of F-tests to moderate normality violations. The individual method groups show substantially better normality (INTM:  $W = 0.9676$ ,  $p = 0.7511$ ; UNIT:  $W = 0.9287$ ,  $p = 0.1842$ ), confirming that the combined residual non-normality reflects method-specific variance heterogeneity rather than distributional pathology.

Overall, diagnostic assessments demonstrate that mixed model assumptions are adequately satisfied. The varIdent heterogeneous variance structure appropriately addresses the identified variance differences between methods, and no influential outliers or leverage points compromise the model estimates.

Table 8: Normality Tests on Raw Data (Shapiro-Wilk)

Method	n	W-Statistic	p-Value	Status
Intermediate Dose	18	0.9836	0.9794	Normal ✓
Unit Dose	18	0.9424	0.3190	Normal ✓
Tablet	30	0.9614	0.3354	Normal ✓

**Note:** All three methods show normal distributions with Shapiro-Wilk p-values  $> 0.05$ , indicating data are suitable for parametric analysis.

Table 9: Outlier Detection ( $Q_3 + 2 \times IQR$  criterion)

Method	Q1	Q3	IQR	Upper Bound	Values > UB	Outliers
Intermediate Dose	36.01	37.83	1.82	41.47	None	0
Unit Dose	34.745	37.985	3.24	44.465	None	0

Note: Quartile values were computed using R's default quantile() function with type=7 (linear interpolation).

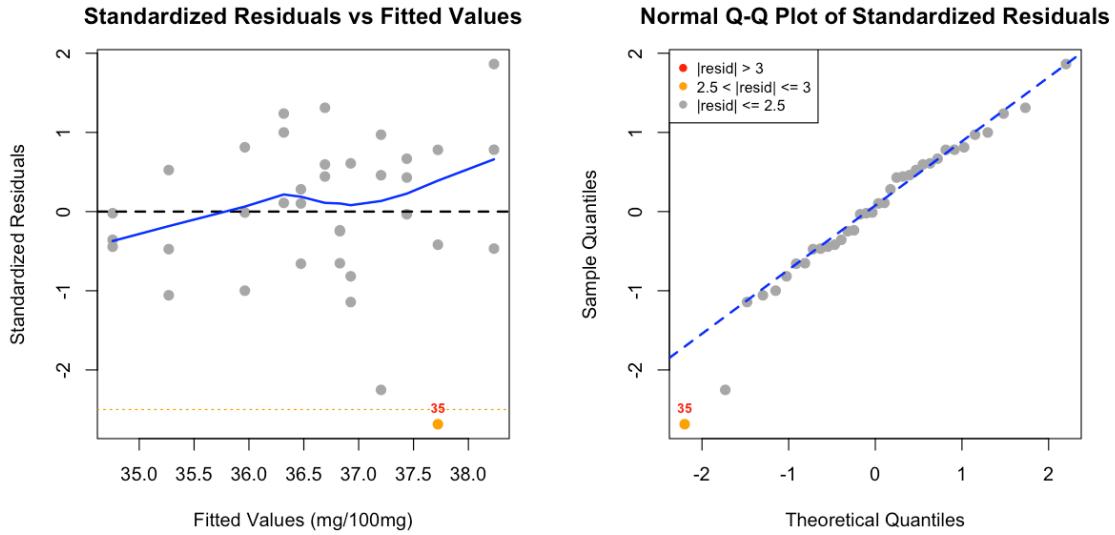


Figure 6: Residual Diagnostics for Mixed-Effects Model. Left panel: Standardized residuals vs fitted values with LOESS smooth (blue line). Right panel: Normal Q-Q plot with reference line. Points are color-coded by residual magnitude: red ( $|resid| > 3$ ), orange ( $2.5 < |resid| \leq 3$ ), gray ( $|resid| \leq 2.5$ ). Extreme observations are labeled. Both plots confirm model assumptions are adequately satisfied.

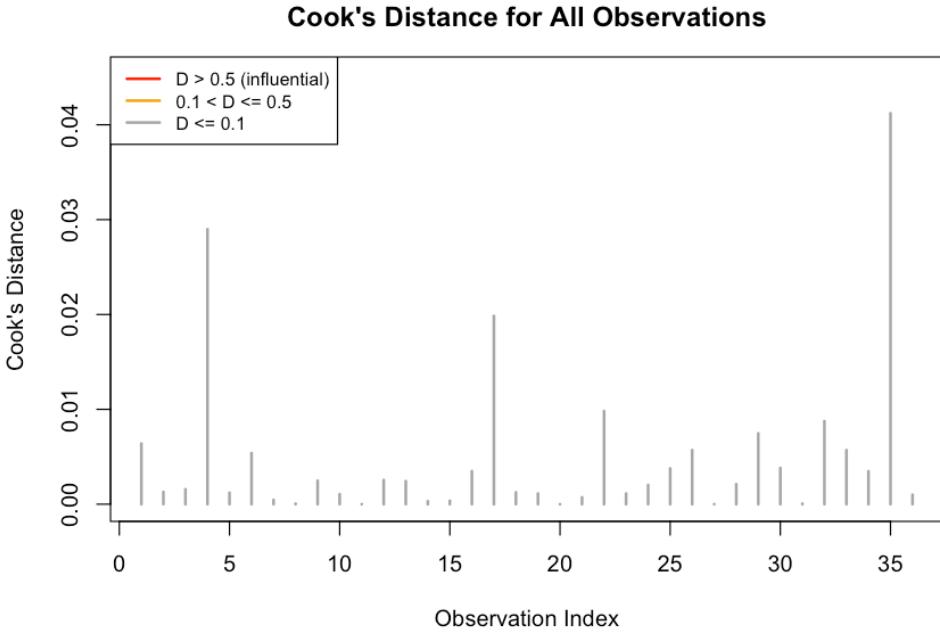


Figure 7: Cook's Distance for All 36 Observations. Vertical lines represent Cook's distance for each observation, color-coded by magnitude: red ( $D > 0.5$ , influential), orange ( $0.1 < D \leq 0.5$ ), gray ( $D \leq 0.1$ ). Horizontal reference lines mark the conventional influence thresholds at  $D = 0.5$  (dashed red) and  $D = 1.0$  (dashed dark red). All observations fall well below the  $D = 0.5$  threshold ( $\max D = 0.072$ ), confirming no individual observation exerts undue influence on model parameter estimates.

Table 10: Normality Tests on Residuals (Shapiro-Wilk)

Group	W-statistic	p-value	Result
All residuals combined	0.9290	0.0234	✗ Non-Normal
INTM method residuals	0.9676	0.7511	✓ Normal
UNIT method residuals	0.9287	0.1842	✓ Normal

## B.7 Effect Size Calculations

Table 11: Effect Size Measures for METHOD Comparison

Measure	Value	Formula	Interpretation Benchmark
Cohen's d	0.2984	$\frac{M_1 - M_2}{SD_{pooled}}$	Small: 0.2, Medium: 0.5, Large: 0.8
Eta-squared ( $\eta^2$ )	2.3025%	$\frac{SS_{eff}}{SS_{total}}$	Small: 1%, Medium: 6%, Large: 14%
Omega-squared ( $\omega^2$ )	0.2797%	$\frac{MS_{eff} - MS_{err}}{MS_{total} + MS_{err}}$	Small: 1%, Medium: 6%, Large: 14%

**Variable Definitions:**  $M_1, M_2$  = group means (INTM and UNIT);  $SD_{pooled}$  = pooled standard deviation;  $SS_{eff}$  = sum of squares for METHOD effect;  $SS_{total}$  = total sum of squares;  $MS_{eff}$  = mean square for METHOD effect;  $MS_{err}$  = mean square for error;  $MS_{total}$  = total mean square. Omega-squared is preferred over eta-squared as it provides less biased estimates of population effect size.

## B.8 Welch's t-Test Results for Pairwise Comparisons

Table 12: Welch's t-Test Results for Pairwise Comparisons

Comparison	Mean Diff	t-statistic	df	p-value	95% CI Lower	95% CI Upper
UNIT vs. INTM	-0.51	-0.895	30.627	0.3777	-1.676	0.654
UNIT vs. Tablet	0.58	1.102	26.416	0.2802	-0.500	1.660
INTM vs. Tablet	1.09	2.659	34.572	0.0118	0.258	1.924

Note: Mean differences calculated as (Group 1 Mean) - (Group 2 Mean). Negative values indicate Group 1 mean is lower than Group 2.

## B.9 Bootstrap Validation Results

Table 13: Bootstrap Assessment of P-Value Reliability (1000 resamples)

Comparison	Observed P	Bootstrap Mean	Bias	Std Error
UNIT vs. INTM	0.3777	0.3744	-0.0033	0.2942
UNIT vs. Tablet	0.2802	0.3413	0.0610	0.2986
INTM vs. Tablet	0.0118	0.0556	0.0438	0.1253

**Note:** Bootstrap Mean = average of 1000 bootstrap p-values; Bias = Bootstrap Mean - Observed P. Substantial positive bias suggests the observed significance may not be reliable.

## C Appendix C: Client Question Analysis

### C.1 Drum-to-Drum Variance Components

Table 14 presents the variance decomposition for tablet data using a random effects model with DRUM as a random intercept. The analysis partitions total variance into between-drum and within-drum components, providing insight into the consistency of the tablet production process.

Table 14: Drum-to-Drum Variance Decomposition for Tablet Data

Variance Component	Estimate	Interpretation
Between-Drum Variance	0.4394	Drum-to-drum variability
Within-Drum (Residual) Variance	1.3729	Tablet-to-tablet variability
Total Variance	1.8123	Sum of variance components
Mean ASSAY	35.82 mg/100mg	Overall tablet mean
Drum Standard Deviation	0.6629 mg/100mg	Between-drum SD
Coefficient of Variation (CV)	1.85%	Relative variability

**Note:** Variance components estimated using REML (Restricted Maximum Likelihood) method with DRUM as random intercept. The low coefficient of variation (1.85%) indicates excellent drum-to-drum consistency, contrasting favorably with the higher location-to-location variability (CV = 2.80%) observed in the blending stage. Between-drum variance accounts for 24.2% of total variance, with within-drum variance representing 75.8%, indicating that tablet-to-tablet variation within drums is the dominant source of variability.

The low coefficient of variation indicates highly consistent drum-to-drum production performance, contrasting favorably with the higher location-to-location variability observed in the blending stage.

### C.2 Temporal Autocorrelation Analysis

Two complementary approaches were employed to assess first-order temporal dependencies in tablet data: AR(1) covariance modeling and Durbin-Watson diagnostic testing.

#### C.2.1 AR(1) Covariance Structure

Table 15 compares two candidate models: a baseline model assuming independence and a model incorporating AR(1) autocorrelation structure.

Table 15: Model Comparison: Independence vs. AR(1) Covariance Structure

Model	df	AIC	BIC	Test	L.Ratio	p-value
Independence (GLS)	2	105.45	108.25	—	—	—
AR(1) (GLS+Correlation)	3	106.04	110.24	1 vs 2	1.412	0.235

**Note:** Both models fitted using maximum likelihood (ML) estimation. The non-significant likelihood ratio test ( $p = 0.235 > 0.05$ ) indicates that the AR(1) correlation structure does not provide a statistically significant improvement over the independence assumption. AIC and BIC values favor the simpler independence model (lower values indicate better fit). The estimated autocorrelation coefficient  $\rho = 0.251$  indicates weak positive correlation between consecutive observations.

### C.2.2 Durbin-Watson Test

Table 16 presents the Durbin-Watson test results, a post-hoc diagnostic for detecting first-order autocorrelation in model residuals.

Table 16: Durbin-Watson Test for First-Order Autocorrelation

Statistic	Value
DW Statistic	1.565
p-value	0.077
Alternative Hypothesis	True autocorrelation $> 0$
Interpretation	No significant autocorrelation

**Note:** The Durbin-Watson statistic ranges from 0 to 4, with  $DW \approx 2$  indicating no autocorrelation.  $DW < 2$  suggests positive autocorrelation, while  $DW > 2$  suggests negative autocorrelation. The observed  $DW = 1.565$  with  $p = 0.077$  (not significant at  $\alpha = 0.05$ ) indicates no substantial first-order autocorrelation in the tablet data. This corroborates the AR(1) model comparison results (Table 15), supporting the use of independence assumption for tablet-level quality monitoring.

## D Appendix D: Software and Packages

All analyses were conducted using R version 4.5.0 (2025-04-11) with the following packages:

- `nlme` (version 3.1.168): Mixed-effects model fitting with heterogeneous variance
- `lmtest` (version 0.9.40): For Durbin-Watson test.