INSTITUT FÜR TESTFORSCHUNG
UND TESTENTWICKLUNG e.V.

**ITT Technical Reports on Language Testing 5**

# Measuring Receptive Lexical Knowledge in Spanish: A Rasch-Model Perspective of the Spanish R-VST 1.2

Erwin Tschirner[1]

[1] Leipzig University, Germany

**Author Note**

Erwin Tschirner https://orcid.org/0000-0002-5915-5344

Correspondence concerning this report should be addressed to Erwin Tschirner, Herder-Institut, Universität Leipzig, Beethovenstraße 15, 04107 Leipzig. email: tschirner@uni-leipzig.de

Bibliographische Informationen der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Die "ITT Technical Reports on Language Testing" sind eine Reihe des Instituts für Testforschung und Testentwicklung e. V. (ITT), in der Technische Reports zu unterschiedlichen Sprachtestevaluationen veröffentlicht werden.

Institut für Testforschung und Testentwicklung e.V. Leipzig
c/o Herder-Institut
Universität Leipzig
Beethovenstraße 15
04107 Leipzig
www.itt-leipzig.de

Herausgeberschaft:
Olaf Bärenfänger, Universität Leipzig
Jupp Möhring, Technische Universität Dresden
Erwin Tschirner, Universität Leipzig

Redaktion:
Lisa Lenort, Elisabeth Muntschick




URN des Bandes: https://nbn-resolving.org/urn:nbn:de:bsz:15-qucosa2-1001355
URN der Reihe: https://nbn-resolving.org/urn:nbn:de:bsz:15-qucosa2-902757

ISSN **2942-741X**

**Table of Contents**



**List of Tables**

**Abstract**

The Spanish Receptive Vocabulary Size Test 1.2 (R-VST) is a revised version of the Spanish R-VST 1, updated in 2023 to improve measurement precision and lexical coverage. Responses from 295 university learners of Spanish (semesters 1–8, plus heritage learners) were analyzed using the Rasch model. The analysis found high person (Rel. = .94) and item (Rel. = .98) reliabilities, indicating that the ranking of persons and items is highly reproducible. The corresponding separation indices (4.00 for persons and 6.90 for items) suggest that the test can distinguish approximately six distinct ability levels among test takers and ten distinct difficulty levels among items. Item difficulties ranged from −3.23 to +2.76 logits and conformed closely to model expectations (Infit $M$ = 0.99, Outfit $M$ = 1.01). Standard errors were small ($M$ = 0.18), and point-measure correlations averaged 0.41, reflecting strong discrimination. ANOVA confirmed that lexical frequency significantly predicted item difficulty, $F(4, 145) = 17.16$, $p < .001$, $\eta^2 = 0.32$. These results demonstrate that the Spanish R-VST 1.2 is a psychometrically robust instrument that validly operationalizes lexical frequency as an index of receptive vocabulary knowledge.

## Introduction

The Spanish Receptive Vocabulary Size Test (R-VST) was originally developed within the ITT Vocabulary Size Test project to provide an empirically grounded measure of lexical comprehension in Spanish. Modeled after Nation's (1990) Vocabulary Levels Test and based on the Routledge Frequency Dictionary of Spanish (Davies & Davies, 2017), the R-VST assesses how many of the most frequent 5,000 Spanish words learners can recognize and understand. It samples ten clusters per frequency band (1,000–5,000 words), each containing six words representing nouns, verbs, and adjectives, yielding 150 multiple-choice items in total. The original R-VST 1, validated in Tschirner (2021), demonstrated high internal consistency (Cronbach's α = .96) and strong construct validity across all five frequency bands.

Version 1.2, introduced in 2023, incorporates revised item wording, optimized distractors, and minor frequency-band adjustments informed by earlier validation studies. These revisions were designed to enhance measurement precision across the full proficiency continuum while maintaining lexical-frequency representativeness and comparability within the ITT framework. The present report documents the psychometric characteristics of the Spanish R-VST 1.2 based on data from 295 university learners of Spanish, focusing on item fit, reliability, and the relationship between lexical frequency and item difficulty as estimated through Rasch modeling.

## Methodology

### Participants

The participants were 295 students enrolled in undergraduate Spanish programs at two U.S. universities. Of these, 163 were from the University of Rhode Island (semesters 1–8) and 132 from Yale University (semesters 1–4 and a beginning heritage-language course). The tests were administered outside of class as part of regular coursework. They were not proctored, but all students signed an honor pledge affirming independent work. The mean oral proficiency of the participants was *Intermediate High* on the ACTFL scale (CEFR B1.2), with a minimum of *Intermediate Low* (A2) and a maximum of *Advanced High* (C1). Their mean reading proficiency was *Intermediate Mid* (A2), ranging from *Novice Low* (below A1) to *Advanced Mid* (B2). (ACTFL, n.d.)

### Instrument

The Spanish Receptive Vocabulary Size Test 1.2 (R-VST 1.2) is a revised version of the Spanish R-VST 1, designed to estimate learners' receptive lexical knowledge of the 5,000 most frequent Spanish words. The test consists of 150 multiple-choice items representing five frequency bands (1,000–5,000 words) with 10 clusters per band. Each cluster contains six words and three synonyms, paraphrases, or gapped sentences (targets). Three of the six words are keys, i. e., they correspond to the three targets, while three words are additional distractors. For each target, the

same six words are presented as multiple-choice options, one of which needs to be selected for each target. Each band, accordingly, consists of 30 items (targets). The maximum score per band is 30, i. e., 3 points per cluster. The maximum composite score for all five bands is 150, i. e., five times 30. Items sample a balance of nouns, verbs, and adjectives across frequency bands, ensuring lexical and morphological diversity. The revision process in 2023 refined item wording and distractors to enhance psychometric precision and coverage across proficiency levels.

### Analysis Framework

Responses were dichotomously scored (1 = correct, 0 = incorrect) and analyzed using the Rasch model to obtain item and person measures on a common logit scale. The analysis was conducted with Winsteps 5.10.2.0 (Linacre, 2025) and supplemented by SPSS for descriptive and ANOVA analyses. Item and person reliabilities, separation indices, fit statistics (Infit/Outfit MNSQ and ZSTD), and standard errors were examined to assess model fit and measurement precision. Acceptable fit was defined as $0.7 \leq \text{MNSQ} \leq 1.3$ for high-stakes applications and 0.5–1.5 for general diagnostic purposes. The relationship between lexical frequency and item difficulty was further examined through a one-way ANOVA of mean item measures across the five frequency bands, followed by Scheffé post-hoc tests. Bootstrapped standard errors were computed to verify precision estimates.

## Results

### Overall Reliability and Fit

To evaluate the psychometric quality of the revised test and to confirm its construct validity, a Rasch analysis was conducted. The Rasch model provides a robust framework for examining item functioning, person ability, and the overall measurement precision of the test on a common interval scale. It allows for direct estimation of item difficulty and person ability in logits, enabling detailed assessment of targeting, reliability, and fit. The analysis yielded a person

reliability of 0.94 and item reliability of 0.98, with corresponding separation indices of 4.0 and 6.9, respectively, indicating that both the test takers and items were well differentiated along the proficiency continuum.

To evaluate unidimensionality, a Principal Components Analysis (PCA) of residuals was conducted. The Rasch measures accounted for 32.9 % of the total variance in the data, which is within the expected range for dichotomously scored language-assessment data (Bond & Fox, 2015; Linacre, 2025). Although this proportion may appear modest, it is typical for Rasch analyses of right/wrong responses, where each item contributes only limited information to the model. The first residual contrast had an eigenvalue of 7.82 (3.5 %), exceeding the nominal 2.0 threshold but accounting for only a small proportion of total variance. Inspection of item loadings revealed 14 items with |loading| ≥ 0.40, distributed evenly across response options (three A-items, five B-items, and six C-items) and frequency bands (1st–5th 1,000-word levels). No consistent pattern in content, position, or difficulty emerged. The contrast therefore appears to represent minor local dependencies rather than a substantive secondary dimension, supporting the conclusion that the test is essentially unidimensional.

These results demonstrate that the R-VST 1.2 provides a stable and finely grained measurement of receptive vocabulary size across a wide range of abilities. Table 1 presents the overall person and item separation reliability and the mean item fit statistics (infit and outfit).

**Table 1**

*Overall Reliability and Fit Statistics for the Spanish R-VST 1.2*

| Measure | Reliability | Separation | Mean Infit (MNSQ) | Mean Infit (ZSTD) | Mean Outfit (MNSQ) | Mean Outfit (ZSTD) |
|---|---|---|---|---|---|---|
| Person | 0.94 | 4.04 | 1.00 | 0.00 | 1.00 | 0.00 |
| Item | 0.98 | 6.90 | 0.99 | 0.03 | 1.01 | 0.10 |

**Item-Level Analysis**

Table 2 summarizes the average item-level fit statistics for the 150 items of the Spanish Receptive Vocabulary Size Test 1.2 administered to 295 Spanish students (s. Appendix A for the complete 150 item-level statistics). Item difficulties ranged from −3.23 (easiest) to +2.76 logits (hardest), with a mean of 0.00 and a standard deviation of 1.28. Item reliability (0.98) and separation (6.87) indicate that the items are well spread across the ability continuum. Following Linacre's (2025) conversion formula for separation strata (4 × separation + 1) / 3, the corresponding separation indices (4.04 for persons and 6.90 for items) suggest that the test can distinguish approximately six distinct ability levels among test takers and ten distinct difficulty levels among items. Average standard errors were small (*M* = .18), demonstrating precise estimation. Infit and outfit mean-square values centered on 1.00 (*M* = .99 and 1.01, respectively), confirming that the data conformed closely to Rasch model expectations. The mean point-measure correlation (0.41, *SD* = 0.10), suggests satisfactory discrimination, while observed exact-match percentages (*M* = 78.5%) aligned closely with expected values (78.8%), further supporting model fit.

**Table 2**

*Average Item-Level Fit Statistics for the 150 Items of the Spanish R-VST 1.2*

| | Total Score | Total Count | Measure | S.E. | Infit MNSQ | ZSTD | Outfit MNSQ | ZSTD | Point-Measure Correlation |
|---|---|---|---|---|---|---|---|---|---|
| Mean | 177.5 | 257.7 | 0.00 | 0.18 | 0.99 | 0.03 | 1.01 | 0.10 | 0.41 |
| S.D. | 60.5 | 25.4 | 1.28 | 0.05 | 0.12 | 1.56 | 0.43 | 1.78 | 0.10 |

Table 2 reports the mean total score the items received (i.e., the number of correct responses); the average total count of examinees who attempted the item; the mean item

difficulty measure in logits; the average standard error of estimation (S.E.); the mean infit and outfit mean-square (MNSQ) fit statistics with their corresponding standardized z-values (ZSTD); and the average point-measure correlation (*item-person correlation*) between each item and the overall test measure. Appendix A provides the item numbers and the corresponding values for all 150 individual items.

The item difficulty measure expresses the relative challenge of each item on the Rasch logit scale, where higher positive values indicate more difficult items and negative values denote easier ones. Item measures in this study ranged from −3.23 to +2.76 logits ($M$ = 0.00, SD = 1.28), confirming broad coverage of the ability continuum.

The standard error (S.E.) reflects the precision of each item's estimated difficulty, with smaller values indicating greater measurement precision (Bond & Fox, 2015; Linacre, 2025). Following common Rasch practice (e.g., Boone et al., 2014), values up to 0.30 are typically interpreted as highly precise, 0.31–0.40 as moderately precise, 0.41–0.50 as imprecise, and values above 0.50 as unstable, suggesting that the item estimate could shift noticeably with additional data.

The bootstrapped mean S.E. was .196 (BCa 95% CI [0.188, 0.204]; observed range = 0.14–0.39; bootstrapped range = 0.16–0.42; $N$ = 150), based on 1,000 resamples, demonstrating excellent precision of item calibration. The five items with the highest standard errors (S.E. = .32–.39) were all located at the lower end of the difficulty continuum (−3.23 ≤ Measure ≤ −2.67 logits). These items were answered correctly by nearly all test takers, resulting in reduced information and correspondingly larger standard errors. Despite this, all five items displayed acceptable fit (Infit and Outfit 0.8–1.3) and positive point-measure correlations, indicating that they functioned as expected at the easy end of the scale.

The infit and outfit mean-square (MNSQ) statistics assess how well each item conforms to Rasch model expectations. Infit is sensitive to responses near a person's ability level, whereas outfit is more influenced by unexpected responses far from that level. Values close to 1.0 indicate excellent fit. MNSQ values between 0.5 and 1.5 are generally acceptable for most purposes, while values between 0.7 and 1.3 are preferred for high-stakes testing. Values below 0.5 suggest redundancy, whereas values above 1.5 indicate noise or unexpected responses, and those exceeding 2.0 are considered unreliable because they may distort measurement (Bond & Fox, 2015; Linacre, 2002; Wright & Linacre, 1994).

In this dataset, the average infit and outfit MNSQ values were 0.99 and 1.01, respectively, indicating excellent overall model fit. The largest infit value (1.44) and the lowest (0.77) confirm very good conformity to model expectations. Three items showed substantial underfit with Outfit MNSQ values above 2.0 (Items @2A = 3.94, @12B = 2.82, and @13B = 2.05). All three had acceptable Infit values (≤ 1.2), suggesting that the misfit was due to a few unexpected responses rather than systematic calibration problems. These items were retained for coverage purposes but will be reviewed in future revisions of the R-VST 1.2.

The standardized fit statistics (ZSTD) provide z-score equivalents of the infit and outfit mean-square values, indicating the statistical significance of any deviation from model expectations. In this dataset, a number of items showed ZSTD values exceeding ±2.0. However, this pattern is expected with large sample sizes, as ZSTD values are highly sensitive to even minor departures from model fit (Linacre, 2002). Following common Rasch practice, item fit was therefore evaluated primarily on the basis of MNSQ values, with ZSTD values interpreted as supplementary indicators. Given that nearly all MNSQ values fell within acceptable limits, the model fit was deemed satisfactory despite the presence of some extreme ZSTD scores.

The point-measure correlation indicates how strongly an item discriminates between higher- and lower-ability examinees. Values above 0.25 are usually regarded as acceptable (Boone, Staver, & Yale, 2014). In the present dataset, the item–person correlations ranged from 0.08 to 0.59 (*M* = 0.41, SD 0.10), indicating generally strong discrimination across the test. Nine items (6%) had correlations below 0.25, suggesting weaker alignment with the overall ability measure. Four of these were extremely easy items (−3.23 ≤ Measure ≤ −2.17 logits) whose limited response variability naturally reduced correlation strength and occasionally inflated outfit values. These items were answered correctly by nearly all examinees, leading to reduced correlations and occasional outlying responses rather than systematic misfit. Three were very difficult items (1.11 ≤ Measure ≤ 2.03 logits) that also yielded lower correlations, most likely because only relatively few examinees responded correctly, limiting the spread of responses. For Item @5C, the moderate difficulty and elevated Outfit (1.64) further suggest that one or more distractors may not have functioned as intended, producing inconsistent response patterns. The remaining two items, @4B and @44A, had moderate difficulty (-0.81 and 0.28 logits, respectively) and modestly elevated outfit statistics, likely reflecting guessing or partial lexical knowledge. Overall, the low-correlation items behaved as expected at the extremes of the ability continuum, and all were retained for content and difficulty coverage.

Taken together, the indices summarized in Table 2 and Appendix A indicate that the Spanish R-VST 1.2 items fit the Rasch model well, were precisely estimated, and displayed strong internal consistency and discrimination across the ability range.

**Lexical Frequency Analysis**

After establishing overall item fit and measurement precision, the relationship between lexical frequency and item difficulty was examined through a one-way ANOVA of the mean item measures across the five frequency bands (Bands 1–5). The analysis revealed a significant effect of

frequency band on item difficulty, $F(4, 145) = 17.16$, $p < .001$, $\eta^2 = .32$, indicating that lexical frequency strongly predicts item difficulty in the R-VST 1.2. Post-hoc Scheffé tests showed a clear and orderly progression in difficulty, with mean measures increasing from Band 1 ($M = -1.02$ logits) to Band 5 ($M = 1.12$ logits). Lower-frequency bands contained significantly more difficult items than higher-frequency bands, while adjacent bands did not differ significantly. These results confirm that the R-VST 1.2 effectively operationalizes lexical frequency as an index of difficulty, producing a well-calibrated hierarchy of receptive vocabulary knowledge. As shown in Table 3, the Scheffé post-hoc comparisons illustrate this gradual increase in item difficulty across the five frequency bands.

**Table 3**

*Post-Hoc Scheffé Test for Differences in Mean Item Difficulty Across Frequency Bands*

| Band | N | Subset 1 | Subset 2 | Subset 3 |
|---|---|---|---|---|
| 1 | 30 | -1.02 | | |
| 2 | 30 | -0.49 | -0.49 | |
| 3 | 30 | | 0.10 | |
| 4 | 30 | | 0.30 | 0.30 |
| 5 | 30 | | | 1.12 |

*Note:* Bands sharing a subset number do not differ significantly at α = .05 (Scheffé test). Mean logit values increase with item difficulty.

## Discussion

The Rasch analysis of the Spanish R-VST 1.2 confirmed that the test provides a reliable and valid measure of receptive vocabulary knowledge across a broad range of proficiency levels. High person (Rel = .94) and item reliabilities (Rel = .98) demonstrate that both the examinees and the items were well separated along the ability continuum, yielding stable and replicable estimates. The narrow spread of standard errors ($M = .18$) indicates high measurement precision across the

scale. Mean infit and outfit values close to 1.00 further support the conclusion that the data fit the Rasch model well.

At the construct level, lexical frequency proved to be a strong predictor of item difficulty. The significant effect observed in the one-way ANOVA, $F(4, 145) = 17.16$, $p < .001$, $\eta^2 = .32$, confirmed that the R-VST 1.2 effectively operationalizes frequency-based lexical knowledge. Item difficulty increased systematically from Band 1 to Band 5, reproducing the expected hierarchy of lexical acquisition and validating the test's sampling design.

At the item level, most items exhibited good fit and satisfactory discrimination ($M = .41$). Items with higher standard errors or slightly inflated outfit values occurred primarily at the extremes of the difficulty range, a common feature of vocabulary tests where very easy or very difficult items provide limited information. A few localized misfits—such as Items @2A, @12B, and @13B—were attributable to random response noise rather than systematic bias. Likewise, several low point-measure correlations reflected ceiling effects for very easy items or the influence of suboptimal distractors for mid-range items (e.g., @5C). These items were retained to preserve lexical and difficulty coverage but will be reviewed in subsequent revisions.

Overall, the findings indicate that the Spanish R-VST 1.2 maintains strong construct alignment and internal consistency while offering improved precision across proficiency levels. The test continues to serve as a psychometrically sound instrument for assessing receptive vocabulary size in Spanish and for investigating lexical development in instructed contexts. Future work should expand the item pool to capture lower-frequency and specialized vocabulary, refine distractors in selected items, and validate the instrument with additional learner populations to further strengthen its generalizability.

## Summary and Conclusion

The Rasch analysis of the Spanish R-VST 1.2 confirmed its strong psychometric quality and construct validity. High person and item reliabilities, small standard errors, and balanced item difficulties demonstrate precise and stable measurement across ability levels. The consistent alignment between lexical frequency and item difficulty supports the test's theoretical foundation and practical interpretability. The Spanish R-VST 1.2 thus provides a reliable and valid measure of receptive vocabulary knowledge for learners of Spanish, suitable for both research and instructional assessment contexts.

**Acknowledgments**

Data collection for this study was made possible through the collaboration of LeAnne Spino (University of Rhode Island) and Jorge Méndez-Seijas (Yale University), as well as the instructors and students who participated in the Spanish programs at both institutions. Their contributions are gratefully acknowledged. Special thanks are also due to the Institute for Test Research and Test Development (ITT e.V., Leipzig) for the development, free provision, and results management of the Spanish placement test used in this project. The author also gratefully acknowledges Jupp Möhring for his insightful and constructive comments on an earlier version of this report.

**References**

ACTFL (n.d.). Assigning CEFR Ratings to ACTFL Assessments. Alexandria, VA: ACTFL. https://www.actfl.org/uploads/files/general/Assigning_CEFR_Ratings_To_ACTFL_Assessments.pdf

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.). Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch analysis in the human sciences (Vol. 10, pp. 978-94). Dordrecht: Springer.

Davies, M., & Davies, K. H. (2017). A Frequency Dictionary of Spanish: Core Vocabulary for Learners. London: Routledge.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions, 16(2), 878. https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2025). Winsteps® Rasch measurement computer program (Version 5.10.2.0) [Computer software]. Winsteps.com.

Nation, I. S. P. (1990). Teaching and learning vocabulary. New York: Newbury House.

Tschirner, E. (2021). Examining the validity and reliability of the ITT Vocabulary Size Tests (*Research Papers in Assessment, 3*). Leipzig: Institut für Testforschung und Testentwicklung.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8(3), 370. https://www.rasch.org/rmt/rmt83b.htm

**Appendix A**

| Item | Total Score | Total Count | Measure | S.E. | Infit MNSQ | ZSTD | Outfit MNSQ | ZSTD | Pt-Measure Correlation |
|---|---|---|---|---|---|---|---|---|---|
| @1A | 265 | 290 | -1.76 | 0.22 | 1.01 | 0.12 | 0.83 | -0.40 | 0.28 |
| @1B | 281 | 292 | -2.74 | 0.32 | 0.89 | -0.33 | 0.80 | -0.28 | 0.27 |
| @1C | 249 | 289 | -1.14 | 0.19 | 0.92 | -0.66 | 0.79 | -0.73 | 0.40 |
| @2A | 274 | 292 | -2.17 | 0.26 | 1.05 | 0.32 | 3.94 | 4.38 | 0.08 |
| @2B | 285 | 292 | -3.23 | 0.39 | 1.01 | 0.15 | 0.89 | -0.02 | 0.16 |
| @2C | 260 | 288 | -1.62 | 0.21 | 0.91 | -0.56 | 0.79 | -0.57 | 0.35 |
| @3A | 267 | 287 | -2.03 | 0.25 | 0.93 | -0.35 | 0.80 | -0.40 | 0.30 |
| @3B | 240 | 283 | -1.03 | 0.18 | 1.03 | 0.34 | 0.82 | -0.65 | 0.35 |
| @3C | 181 | 280 | 0.35 | 0.14 | 1.17 | 2.71 | 1.71 | 4.43 | 0.31 |
| @4A | 259 | 283 | -1.77 | 0.23 | 0.89 | -0.66 | 0.57 | -1.29 | 0.37 |
| @4B | 236 | 286 | -0.81 | 0.17 | 1.16 | 1.53 | 1.94 | 3.24 | 0.21 |
| @4C | 220 | 281 | -0.49 | 0.16 | 1.02 | 0.24 | 0.95 | -0.21 | 0.39 |
| @5A | 263 | 283 | -1.99 | 0.24 | 0.89 | -0.56 | 0.51 | -1.42 | 0.35 |
| @5B | 181 | 275 | 0.31 | 0.14 | 0.89 | -1.87 | 0.76 | -1.82 | 0.53 |
| @5C | 136 | 269 | 1.11 | 0.14 | 1.33 | 5.33 | 1.64 | 5.30 | 0.23 |
| @6A | 206 | 275 | -0.27 | 0.15 | 0.99 | -0.16 | 0.90 | -0.52 | 0.43 |
| @6B | 141 | 279 | 1.11 | 0.14 | 0.87 | -2.39 | 0.91 | -0.87 | 0.56 |
| @6C | 253 | 283 | -1.50 | 0.21 | 0.99 | -0.04 | 0.71 | -0.90 | 0.34 |
| @7A | 157 | 275 | 0.76 | 0.14 | 1.05 | 0.92 | 1.00 | 0.06 | 0.44 |
| @7B | 228 | 281 | -0.72 | 0.17 | 0.99 | -0.03 | 0.78 | -0.96 | 0.40 |
| @7C | 195 | 276 | 0.00 | 0.15 | 0.94 | -0.88 | 0.82 | -1.14 | 0.48 |
| @8A | 167 | 272 | 0.54 | 0.14 | 1.00 | 0.05 | 0.90 | -0.76 | 0.47 |
| @8B | 263 | 284 | -1.94 | 0.24 | 0.91 | -0.48 | 0.55 | -1.29 | 0.35 |
| @8C | 179 | 272 | 0.29 | 0.14 | 1.06 | 1.05 | 1.03 | 0.25 | 0.41 |
| @9A | 228 | 280 | -0.74 | 0.17 | 0.99 | -0.05 | 0.86 | -0.54 | 0.39 |
| @9B | 250 | 283 | -1.39 | 0.20 | 0.99 | -0.01 | 1.64 | 1.85 | 0.30 |
| @9C | 277 | 286 | -2.93 | 0.35 | 0.95 | -0.06 | 1.76 | 1.35 | 0.18 |
| @10A | 248 | 283 | -1.27 | 0.19 | 1.05 | 0.44 | 1.22 | 0.79 | 0.28 |
| @10B | 233 | 280 | -0.86 | 0.17 | 0.87 | -1.31 | 0.69 | -1.34 | 0.45 |
| @10C | 274 | 285 | -2.67 | 0.32 | 0.84 | -0.56 | 0.46 | -1.23 | 0.32 |
| @11A | 237 | 282 | -0.94 | 0.18 | 0.95 | -0.43 | 0.79 | -0.83 | 0.40 |
| @11B | 284 | 291 | -3.23 | 0.39 | 0.91 | -0.16 | 0.57 | -0.72 | 0.24 |
| @11C | 242 | 286 | -1.01 | 0.18 | 1.00 | 0.00 | 0.82 | -0.67 | 0.37 |
| @12A | 262 | 285 | -1.84 | 0.23 | 0.90 | -0.58 | 0.65 | -0.95 | 0.35 |
| @12B | 221 | 282 | -0.50 | 0.16 | 1.01 | 0.11 | 2.82 | 6.16 | 0.35 |
| @12C | 242 | 281 | -1.12 | 0.19 | 0.94 | -0.51 | 0.74 | -0.95 | 0.39 |
| @13A | 254 | 282 | -1.60 | 0.21 | 0.83 | -1.18 | 0.54 | -1.58 | 0.42 |

| Item | Total Score | Total Count | Measure | S.E. | Infit MNSQ | ZSTD | Outfit MNSQ | ZSTD | Pt-Measure Correlation |
|---|---|---|---|---|---|---|---|---|---|
| @13B | 185 | 279 | 0.24 | 0.14 | 1.18 | 2.85 | 2.05 | 5.82 | 0.25 |
| @13C | 206 | 283 | -0.14 | 0.15 | 0.89 | -1.64 | 0.76 | -1.46 | 0.51 |
| @14A | 230 | 280 | -0.79 | 0.17 | 1.01 | 0.13 | 0.95 | -0.13 | 0.37 |
| @14B | 279 | 287 | -3.05 | 0.37 | 0.86 | -0.34 | 0.37 | -1.38 | 0.29 |
| @14C | 225 | 279 | -0.67 | 0.17 | 0.85 | -1.62 | 0.63 | -1.87 | 0.49 |
| @15A | 228 | 275 | -0.86 | 0.18 | 0.96 | -0.34 | 0.77 | -0.94 | 0.40 |
| @15B | 260 | 284 | -1.78 | 0.23 | 1.03 | 0.24 | 1.25 | 0.76 | 0.25 |
| @15C | 233 | 280 | -0.87 | 0.17 | 0.95 | -0.41 | 0.93 | -0.19 | 0.39 |
| @16A | 242 | 275 | -1.33 | 0.20 | 0.95 | -0.31 | 1.37 | 1.19 | 0.31 |
| @16B | 186 | 262 | 0.02 | 0.15 | 0.84 | -2.40 | 0.70 | -1.98 | 0.54 |
| @16C | 103 | 265 | 1.77 | 0.15 | 1.03 | 0.51 | 1.03 | 0.32 | 0.47 |
| @17A | 213 | 274 | -0.42 | 0.16 | 0.93 | -0.84 | 0.81 | -0.95 | 0.45 |
| @17B | 215 | 274 | -0.50 | 0.16 | 0.96 | -0.46 | 0.82 | -0.86 | 0.43 |
| @17C | 212 | 271 | -0.46 | 0.16 | 0.91 | -1.10 | 0.73 | -1.41 | 0.47 |
| @18A | 182 | 256 | -0.01 | 0.15 | 0.90 | -1.45 | 0.78 | -1.28 | 0.50 |
| @18B | 134 | 259 | 1.04 | 0.14 | 0.88 | -2.20 | 0.81 | -1.85 | 0.57 |
| @18C | 189 | 270 | 0.06 | 0.15 | 0.98 | -0.28 | 0.85 | -0.90 | 0.46 |
| @19A | 147 | 263 | 0.84 | 0.14 | 1.00 | -0.01 | 0.95 | -0.43 | 0.48 |
| @19B | 198 | 262 | -0.28 | 0.16 | 1.00 | 0.06 | 0.87 | -0.64 | 0.42 |
| @19C | 89 | 257 | 1.99 | 0.15 | 1.19 | 2.54 | 1.40 | 2.88 | 0.35 |
| @20A | 142 | 273 | 1.06 | 0.14 | 0.99 | -0.24 | 0.95 | -0.47 | 0.49 |
| @20B | 173 | 279 | 0.54 | 0.14 | 1.13 | 2.20 | 1.26 | 2.00 | 0.35 |
| @20C | 231 | 273 | -0.97 | 0.18 | 0.85 | -1.42 | 0.61 | -1.67 | 0.47 |
| @21A | 135 | 275 | 1.19 | 0.14 | 0.96 | -0.62 | 1.05 | 0.52 | 0.49 |
| @21B | 138 | 272 | 1.10 | 0.14 | 1.06 | 1.13 | 1.17 | 1.56 | 0.43 |
| @21C | 271 | 288 | -2.21 | 0.26 | 0.97 | -0.10 | 0.56 | -1.11 | 0.30 |
| @22A | 142 | 259 | 0.89 | 0.14 | 0.93 | -1.27 | 0.88 | -1.03 | 0.53 |
| @22B | 223 | 269 | -0.84 | 0.18 | 0.84 | -1.59 | 0.64 | -1.59 | 0.48 |
| @22C | 249 | 279 | -1.49 | 0.21 | 0.77 | -1.72 | 0.45 | -2.03 | 0.47 |
| @23A | 191 | 278 | 0.13 | 0.15 | 1.06 | 1.00 | 1.05 | 0.37 | 0.40 |
| @23B | 130 | 272 | 1.25 | 0.14 | 1.09 | 1.47 | 1.26 | 2.41 | 0.41 |
| @23C | 230 | 275 | -0.90 | 0.18 | 0.82 | -1.75 | 0.73 | -1.12 | 0.48 |
| @24A | 121 | 253 | 1.21 | 0.14 | 1.17 | 2.71 | 1.21 | 1.86 | 0.38 |
| @24B | 229 | 273 | -0.95 | 0.18 | 0.90 | -0.95 | 0.66 | -1.40 | 0.44 |
| @24C | 229 | 275 | -0.87 | 0.18 | 0.97 | -0.23 | 1.10 | 0.47 | 0.36 |
| @25A | 198 | 264 | -0.28 | 0.16 | 0.98 | -0.25 | 0.84 | -0.80 | 0.44 |
| @25B | 170 | 264 | 0.39 | 0.15 | 0.90 | -1.71 | 0.79 | -1.56 | 0.53 |
| @25C | 234 | 276 | -1.01 | 0.18 | 0.94 | -0.48 | 0.80 | -0.73 | 0.40 |

| Item | Total Score | Total Count | Measure | S.E. | Infit MNSQ | ZSTD | Outfit MNSQ | ZSTD | Pt-Measure Correlation |
|---|---|---|---|---|---|---|---|---|---|
| @26A | 164 | 254 | 0.34 | 0.15 | 1.11 | 1.69 | 1.12 | 0.87 | 0.39 |
| @26B | 237 | 274 | -1.18 | 0.19 | 0.87 | -1.10 | 0.74 | -0.90 | 0.42 |
| @26C | 147 | 253 | 0.70 | 0.15 | 0.90 | -1.73 | 0.83 | -1.43 | 0.54 |
| @27A | 106 | 238 | 1.42 | 0.15 | 1.12 | 1.83 | 1.28 | 2.28 | 0.40 |
| @27B | 138 | 247 | 0.81 | 0.15 | 1.16 | 2.53 | 1.29 | 2.25 | 0.37 |
| @27C | 115 | 249 | 1.34 | 0.15 | 1.03 | 0.51 | 1.14 | 1.31 | 0.45 |
| @28A | 135 | 245 | 0.89 | 0.15 | 0.88 | -2.16 | 0.76 | -2.19 | 0.58 |
| @28B | 192 | 259 | -0.19 | 0.16 | 0.90 | -1.34 | 0.86 | -0.76 | 0.49 |
| @28C | 165 | 259 | 0.43 | 0.15 | 0.90 | -1.62 | 0.82 | -1.39 | 0.53 |
| @29A | 192 | 251 | -0.36 | 0.17 | 0.93 | -0.82 | 0.79 | -1.03 | 0.46 |
| @29B | 122 | 245 | 1.16 | 0.15 | 1.03 | 0.48 | 1.13 | 1.21 | 0.45 |
| @29C | 228 | 265 | -1.13 | 0.19 | 1.02 | 0.18 | 1.03 | 0.20 | 0.33 |
| @30A | 160 | 256 | 0.48 | 0.15 | 1.03 | 0.44 | 0.99 | 0.00 | 0.44 |
| @30B | 171 | 263 | 0.35 | 0.15 | 1.05 | 0.86 | 0.99 | -0.04 | 0.42 |
| @30C | 173 | 262 | 0.29 | 0.15 | 0.85 | -2.42 | 0.73 | -2.00 | 0.55 |
| @31A | 169 | 256 | 0.31 | 0.15 | 0.91 | -1.48 | 0.80 | -1.42 | 0.53 |
| @31B | 184 | 261 | 0.01 | 0.15 | 0.96 | -0.59 | 0.84 | -0.95 | 0.47 |
| @31C | 178 | 266 | 0.25 | 0.15 | 0.91 | -1.46 | 0.79 | -1.46 | 0.52 |
| @32A | 233 | 269 | -1.20 | 0.19 | 0.91 | -0.70 | 0.81 | -0.59 | 0.40 |
| @32B | 172 | 255 | 0.21 | 0.15 | 0.93 | -1.11 | 0.77 | -1.55 | 0.51 |
| @32C | 96 | 237 | 1.61 | 0.15 | 1.07 | 1.10 | 1.22 | 1.83 | 0.43 |
| @33A | 77 | 229 | 2.04 | 0.16 | 1.32 | 3.72 | 1.90 | 5.18 | 0.25 |
| @33B | 119 | 240 | 1.20 | 0.15 | 0.90 | -1.70 | 0.84 | -1.49 | 0.55 |
| @33C | 219 | 262 | -0.89 | 0.18 | 0.96 | -0.35 | 0.78 | -0.85 | 0.40 |
| @34A | 168 | 257 | 0.33 | 0.15 | 0.91 | -1.52 | 0.77 | -1.68 | 0.52 |
| @34B | 236 | 272 | -1.20 | 0.19 | 0.87 | -1.03 | 0.85 | -0.43 | 0.42 |
| @34C | 191 | 258 | -0.20 | 0.16 | 0.80 | -2.70 | 0.62 | -2.30 | 0.56 |
| @35A | 141 | 246 | 0.72 | 0.15 | 1.22 | 3.51 | 1.31 | 2.33 | 0.32 |
| @35B | 147 | 244 | 0.59 | 0.15 | 0.86 | -2.44 | 0.73 | -2.21 | 0.57 |
| @35C | 237 | 263 | -1.59 | 0.22 | 0.82 | -1.17 | 0.53 | -1.52 | 0.43 |
| @36A | 240 | 266 | -1.62 | 0.22 | 0.88 | -0.76 | 0.53 | -1.48 | 0.40 |
| @36B | 224 | 254 | -1.34 | 0.21 | 0.94 | -0.42 | 0.72 | -0.88 | 0.38 |
| @36C | 121 | 240 | 1.13 | 0.15 | 0.99 | -0.16 | 0.96 | -0.33 | 0.49 |
| @37A | 129 | 228 | 0.80 | 0.15 | 1.11 | 1.84 | 1.19 | 1.46 | 0.39 |
| @37B | 136 | 237 | 0.79 | 0.15 | 0.93 | -1.22 | 0.87 | -1.10 | 0.53 |
| @37C | 109 | 229 | 1.30 | 0.15 | 0.90 | -1.67 | 0.99 | -0.07 | 0.55 |
| @38A | 191 | 250 | -0.31 | 0.17 | 0.96 | -0.50 | 0.79 | -1.08 | 0.45 |
| @38B | 77 | 212 | 1.90 | 0.17 | 0.95 | -0.60 | 1.01 | 0.10 | 0.52 |

| Item | Total Score | Total Count | Measure | S.E. | Infit MNSQ | ZSTD | Outfit MNSQ | ZSTD | Pt-Measure Correlation |
|---|---|---|---|---|---|---|---|---|---|
| @38C | 113 | 223 | 1.16 | 0.15 | 1.04 | 0.67 | 1.12 | 1.02 | 0.45 |
| @39A | 165 | 253 | 0.30 | 0.15 | 1.05 | 0.76 | 1.05 | 0.39 | 0.41 |
| @39B | 167 | 242 | 0.13 | 0.16 | 0.94 | -0.90 | 0.83 | -1.04 | 0.49 |
| @39C | 182 | 249 | -0.14 | 0.16 | 0.92 | -1.08 | 0.77 | -1.31 | 0.49 |
| @40A | 155 | 228 | 0.16 | 0.16 | 1.07 | 0.96 | 0.98 | -0.05 | 0.41 |
| @40B | 149 | 237 | 0.44 | 0.15 | 1.04 | 0.65 | 0.98 | -0.07 | 0.44 |
| @40C | 72 | 211 | 2.02 | 0.17 | 1.05 | 0.62 | 1.42 | 2.64 | 0.44 |
| @41A | 95 | 216 | 1.50 | 0.16 | 0.97 | -0.39 | 1.00 | 0.01 | 0.53 |
| @41B | 127 | 253 | 1.15 | 0.15 | 1.08 | 1.26 | 1.22 | 1.96 | 0.43 |
| @41C | 133 | 241 | 0.85 | 0.15 | 0.96 | -0.61 | 0.93 | -0.59 | 0.51 |
| @42A | 129 | 229 | 0.80 | 0.15 | 1.10 | 1.56 | 1.05 | 0.41 | 0.43 |
| @42B | 143 | 236 | 0.56 | 0.15 | 0.94 | -0.97 | 0.87 | -0.95 | 0.52 |
| @42C | 108 | 228 | 1.27 | 0.15 | 0.85 | -2.46 | 0.80 | -1.81 | 0.59 |
| @43A | 72 | 225 | 2.08 | 0.17 | 0.89 | -1.43 | 1.01 | 0.10 | 0.54 |
| @43B | 160 | 234 | 0.13 | 0.16 | 0.98 | -0.22 | 0.87 | -0.74 | 0.46 |
| @43C | 76 | 227 | 2.03 | 0.16 | 1.39 | 4.38 | 1.55 | 3.42 | 0.24 |
| @44A | 156 | 236 | 0.28 | 0.15 | 1.30 | 4.31 | 1.63 | 3.50 | 0.20 |
| @44B | 142 | 218 | 0.31 | 0.16 | 0.92 | -1.27 | 0.85 | -0.92 | 0.51 |
| @44C | 125 | 228 | 0.89 | 0.15 | 1.03 | 0.51 | 1.02 | 0.16 | 0.46 |
| @45A | 159 | 226 | 0.04 | 0.16 | 1.10 | 1.36 | 1.12 | 0.71 | 0.37 |
| @45B | 86 | 232 | 1.84 | 0.16 | 1.12 | 1.61 | 1.21 | 1.59 | 0.40 |
| @45C | 102 | 215 | 1.28 | 0.16 | 1.07 | 0.99 | 1.20 | 1.58 | 0.44 |
| @46A | 70 | 204 | 1.97 | 0.17 | 1.21 | 2.38 | 1.51 | 3.07 | 0.35 |
| @46B | 74 | 208 | 1.93 | 0.17 | 1.44 | 4.74 | 1.59 | 3.58 | 0.23 |
| @46C | 44 | 194 | 2.76 | 0.20 | 1.19 | 1.63 | 1.52 | 2.07 | 0.35 |
| @47A | 140 | 209 | 0.20 | 0.17 | 0.89 | -1.59 | 0.81 | -1.11 | 0.53 |
| @47B | 149 | 221 | 0.21 | 0.16 | 0.91 | -1.35 | 0.75 | -1.53 | 0.52 |
| @47C | 124 | 214 | 0.72 | 0.16 | 0.90 | -1.55 | 0.77 | -1.75 | 0.55 |
| @48A | 68 | 195 | 1.91 | 0.18 | 1.36 | 3.83 | 1.82 | 4.49 | 0.25 |
| @48B | 52 | 196 | 2.44 | 0.19 | 1.24 | 2.25 | 1.51 | 2.39 | 0.33 |
| @48C | 111 | 200 | 0.85 | 0.16 | 1.16 | 2.31 | 1.19 | 1.33 | 0.38 |
| @49A | 151 | 222 | 0.15 | 0.16 | 1.02 | 0.27 | 0.93 | -0.35 | 0.44 |
| @49B | 54 | 213 | 2.56 | 0.18 | 1.24 | 2.24 | 1.46 | 2.22 | 0.32 |
| @49C | 76 | 208 | 1.85 | 0.17 | 1.06 | 0.76 | 1.11 | 0.87 | 0.45 |
| @50A | 204 | 239 | -1.10 | 0.20 | 0.93 | -0.55 | 1.81 | 2.32 | 0.35 |
| @50B | 87 | 207 | 1.50 | 0.16 | 1.12 | 1.60 | 1.19 | 1.50 | 0.42 |
| @50C | 136 | 228 | 0.61 | 0.15 | 1.19 | 2.96 | 1.41 | 2.76 | 0.31 |