Journal of Clinical Medicine

MDPI

Article

Seven-Day Changes in Benton Visual Retention Test-Copying Performance and Qualitative Drawing Observations in Schizophrenia and Acute and Transient Psychotic Disorders: An Observational Study

Violeta Mancas [1], Andrei Gabriel Zanfir [1,*] and Simona Trifu [2,*]

[1] Doctoral School, Carol Davila University of Medicine and Pharmacy, 020021 Bucharest, Romania; violeta.mancas@drd.umfcd.ro
[2] Department of Clinical Neurosciences, Carol Davila University of Medicine and Pharmacy, 020021 Bucharest, Romania
* Correspondence: andrei-gabriel.zanfir@drd.umfcd.ro (A.G.Z.); simona.trifu@umfcd.ro (S.T.)

Abstract

Objective: To estimate seven-day change in Benton Visual Retention Test (BVRT) copying performance and examine associations with baseline negative-subscale scores on the Positive and Negative Syndrome Scale (PANSS). **Methods**: Consecutive recruitment on two psychiatric wards yielded 200 adults with paired assessments from 248 screened patients: 113 with schizophrenia and 87 with acute and transient psychotic disorders (ICD-10 F23). The same ten BVRT Form C cards were used for Administration C (copying) seven days apart. The outcome was change in correctly copied cards (0–10). Baseline PANSS associations were examined by Spearman correlation and regression of T2 adjusted for T1, age, sex, education and diagnosis, with HC3 standard errors. Post hoc analyses included an extended multivariable model incorporating all three PANSS subscales and a Poisson generalized estimating equations analysis of time-by-category differences in qualitative drawing observations. **Results**: Mean BVRT copying scores increased from 6.52 (SD 2.90) to 7.36 (SD 2.84), corresponding to a mean change of 0.835 points (95% CI 0.618–1.052; $p < 0.001$). Scores increased in 120 participants (60.0%). Baseline PANSS N was inversely correlated with BVRT T1 (Spearman rho = −0.313) and remained associated with baseline copying performance after simultaneous adjustment for PANSS P, PANSS G, and demographic and diagnostic covariates (B = −0.170; 95% CI −0.301 to −0.039; $p = 0.011$). Its additional association with BVRT T2 after baseline adjustment remained uncertain (B = −0.031; 95% CI −0.087 to 0.026; $p = 0.286$). Recorded qualitative drawing observations decreased, but total counts closely mirrored correct-card scores (Spearman $\varrho = -0.965$), and evidence that changes differed between observation categories was weak (Wald $\chi^2(5) = 11.27$; $p = 0.046$; cluster-bootstrap $p = 0.065$). Diagnostic-group comparisons remained inconclusive. **Conclusions**: Seven-day same-form BVRT copying retesting was associated with increased accuracy fewer recorded qualitative drawing observations; differences between observation categories were uncertain. Negative symptom severity was associated with poorer concurrent copying performance, without clear evidence of an additional association with retest scores after baseline adjustment. These findings provide descriptive information on short-interval visuoconstructive performance in hospitalized psychotic disorders but do not establish treatment efficacy, improved visual memory, or clinically meaningful individual recovery.

Received: 24 October 2026
Revised: 13 November 2026
Accepted: 16 November 2026
Published: 23 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: Benton Visual Retention Test; copying; schizophrenia; acute and transient psychotic disorders; PANSS; repeated assessment

1. Introduction

Cognitive impairments represent a central feature of schizophrenia and other psychotic disorders, affecting multiple domains, including attention, processing speed, working memory, executive functioning, and visuospatial abilities [1]. The MATRICS Consensus Cognitive Battery was developed to provide standardized assessment across distinct cognitive domains [2,3]. However, the interpretation of individual neuropsychological tests requires careful consideration of the specific cognitive processes involved [4]. In Administration C of the Benton Visual Retention Test (BVRT), the stimulus remains visible throughout the drawing task, and performance therefore primarily reflects visuoperceptual, visuoconstructive, attentional, and motor demands rather than visual memory retention [5,6]. Previous research has emphasized the importance of administration procedures and demographic characteristics in interpreting BVRT performance [7,8]. Factor-analytic evidence has also supported the distinction between copying and immediate visual-memory abilities [9,10].

Although the number of correctly copied BVRT cards provides a convenient quantitative measure of BVRT performance [11], qualitative characteristics of drawing responses may offer complementary descriptive information [12]. BVRT assessment traditionally recognizes several error categories, including omissions, distortions, perseverations, rotations, misplacements, and size errors [11]. Previous longitudinal investigations have demonstrated that specific BVRT error categories may exhibit different patterns of change, particularly in aging populations [8,13]. More recently, qualitative abnormalities involving spatial organization, distortions, perseverations, and misplacements have been reported in patients with schizophrenia in cross-sectional comparisons with substance-induced psychotic disorders [11]. Nevertheless, these findings arise from different populations, administration procedures, and observation intervals. Although previous studies have examined BVRT error patterns in schizophrenia [11] and changes in drawing-based cognitive performance during psychiatric hospitalization [14] the short-term evolution of qualitative drawing characteristics during repeated BVRT copying assessment in hospitalized patients with psychotic disorders remains less well characterized [11,14]. This gap may reflect the predominance of cross-sectional investigations of BVRT error patterns [11] and the use of alternative drawing-based instruments (such as the Clock Drawing Test), rather than repeated BVRT copying assessments, in studies examining cognitive changes during psychiatric hospitalization [14]. Furthermore, qualitative observations and binary correct-card scores represent distinct levels of description and should not be considered interchangeable, as the former characterize specific drawing abnormalities, whereas the latter summarize overall reproduction accuracy [11,15].

Psychopathological symptom dimensions may provide an additional context for understanding differences in visuoconstructive performance. Negative symptoms, as assessed using the Positive and Negative Syndrome Scale (PANSS), have been associated with cognitive impairment across psychotic disorders [16,17]. However, the conventional PANSS negative subscale includes items reflecting cognitive and disorganization-related phenomena, particularly difficulty in abstract thinking (N5) and stereotyped thinking (N7), which are no longer considered core negative symptoms within contemporary conceptual and factor-analytic frameworks [18–20]. Contemporary analyses of the PANSS dimensional structure indicate that difficulty in abstract thinking (N5) clusters within the cognitive/disorganized dimension alongside conceptual disorganization (P2) and poor

https://doi.org/10.3390/jcm15239201

attention (G11), while stereotyped thinking (N7) is also associated with cognitive/disorganized symptoms rather than the core negative symptom dimension [21,22]. These items share cognitive demands with visuoconstructive copying, such as abstraction, flexible organization of the figure and planning, so part of the observed association may reflect cognitive and disorganization features rather than diminished expression or motivation. Consequently, associations between PANSS N scores and neuropsychological performance require cautious interpretation. In addition, positive symptoms and general psychopathology may covary with negative symptom severity [23], making multivariable assessment relevant when examining the specific relationship between conventional PANSS negative-subscale scores and copying performance [24].

Repeated neuropsychological assessment introduces further interpretative challenges. Practice effects, including familiarity with stimulus material, acquired task strategies, and procedural learning, may contribute to improved performance independently of genuine cognitive recovery [25,26]. These effects have been extensively discussed in schizophrenia research and may vary according to the cognitive domain, assessment interval, and use of identical or alternative test forms [27–29]. During acute psychiatric hospitalization, repeated assessment occurs alongside pharmacological treatment, changes in psychopathological severity, and possible variations in attention, motivation, and psychomotor functioning [14,30,31]. Consequently, short-term improvement in neuropsychological test scores cannot automatically be attributed to treatment efficacy or clinical stabilization [26]. Furthermore, the bounded nature of the BVRT correct-card score may introduce ceiling effects, limiting the potential for measurable improvement among participants with high initial performance [32,33].

The neurobiological substrates of cognitive impairment vary across psychiatric and neurological disorders, reflecting heterogeneous molecular, synaptic, and neural-network alterations [34,35]. For example, studies of dementia have documented associations between cognitive deterioration and structural alterations involving the hippocampus, cerebral white matter, and cerebrovascular systems [36,37]. Although these findings concern different clinical populations and cannot be directly extrapolated to short-term changes in psychotic disorders, they emphasize the importance of distinguishing observable cognitive performance from its potential underlying neurobiological mechanisms. This distinction is particularly relevant to the Benton Visual Retention Test (BVRT), as different administration procedures engage different cognitive processes. In Administration C, visual stimuli remain available throughout reproduction, making copying performance primarily dependent on visuoperceptual, visuoconstructive, attentional, and motor processes rather than visual memory retention [38,39]. Accordingly, the present study focuses on the number of correctly copied cards and separately coded qualitative drawing observations as two behavioral descriptions of visuoconstructive performance. Their evaluation over a seven-day interval allows the characterization of short-term changes in copying accuracy and drawing organization without assuming that these changes reflect underlying structural brain modifications, enhanced visual memory, or neurobiological recovery.

The comparison between schizophrenia and acute and transient psychotic disorders (ATPD; ICD-10 F23) is also clinically relevant. Although these diagnostic categories differ in their typical clinical presentation and course, their cognitive profiles may overlap, as previous longitudinal evidence has identified differences in psychopathological severity and functioning without corresponding significant differences in core cognitive domains between schizophrenia and acute and transient psychotic disorders [40]. A longitudinal investigation comparing first-episode ATPD and schizophrenia reported no statistically significant differences in the trajectories of several cognitive domains over a 12-month period, despite differences in clinical symptomatology and functioning [40]. However, these findings cannot be directly generalized to short-interval visuoconstructive copying

performance during hospitalization. Moreover, ATPD encompasses heterogeneous clinical presentations and subsequent diagnostic trajectories [41,42]. Further characterization of repeated copying performance in patients with these diagnoses may therefore contribute descriptive information without presupposing distinct cognitive recovery patterns.

Although associations between psychopathology and neuropsychological performance [24,26], as well as practice-related changes in schizophrenia [29], have been extensively investigated, evidence remains more limited regarding the short-interval evolution of BVRT copying accuracy and separately coded qualitative drawing observations during acute psychiatric hospitalization. Characterizing these outcomes together may offer a more detailed description of repeated visuoconstructive performance while avoiding assumptions about treatment-related cognitive recovery.

The primary objective of the present observational study was to estimate the seven-day change in the number of correctly copied BVRT cards among hospitalized patients with schizophrenia or ATPD. Secondary objectives examined the association between baseline PANSS negative-subscale severity and concurrent copying performance, as well as its additional association with retest performance after adjustment for baseline BVRT scores and relevant covariates. Diagnostic comparisons and qualitative drawing observations were considered exploratory outcomes. Additional post hoc analyses examined the longitudinal heterogeneity of qualitative drawing-observation categories and the concurrent PANSS N association after accounting for positive and general psychopathology dimensions. The F23 group was defined according to ICD-10 diagnostic criteria and was not restricted to first-episode psychosis [43,44]. By distinguishing quantitative copying accuracy from separately coded qualitative observations, the study aimed to provide a more detailed descriptive characterization of short-interval repeated visuoconstructive performance in hospitalized patients with psychotic disorders.

Based on previous evidence regarding repeated neuropsychological assessment and the relationship between psychopathological symptoms and cognitive performance, the present study examined two main hypotheses. First, we hypothesized that visuoconstructive copying performance, measured by the number of correctly copied Benton Visual Retention Test (BVRT) cards, would improve over the seven-day assessment interval in hospitalized patients with schizophrenia or acute and transient psychotic disorders. Second, we hypothesized that greater baseline negative symptom severity, as measured by the PANSS negative subscale, would be associated with poorer concurrent BVRT copying performance. Additionally, we investigated whether baseline PANSS negative-subscale scores were associated with retest performance independently of initial copying accuracy and relevant demographic and diagnostic characteristics. Potential differences between diagnostic groups and changes in qualitative drawing-observation categories were examined on an exploratory basis, without assuming specific directional effects. Given the observational design and repeated administration of identical BVRT stimuli, improvements in copying performance were not presumed to represent treatment-related cognitive recovery or enhanced visual memory.

2. Materials and Methods

2.1. Design and Participants

This observational study was conducted in two psychiatric wards operating in parallel, one male and one female, at the Prof. Dr. Alexandru Obregia Clinical Hospital of Psychiatry, Bucharest, Romania, between 5 November 2024 and 1 October 2026. Participant recruitment was coordinated and conducted by a single psychiatrist across both wards. Potentially eligible patients were identified through consecutive screening of hospital admissions. The psychiatrist was responsible for verifying eligibility according to the established inclusion and exclusion criteria, explaining the study objectives and

assessment procedures, and obtaining written informed consent before study-related evaluations.

Written informed consent was obtained by the psychiatrist responsible for study recruitment before any study-specific assessments. Given the acute psychiatric setting, the psychiatrist evaluated each patient's ability to understand the purpose and procedures of the study, comprehend the information provided, and communicate a voluntary decision regarding participation. This assessment was conducted through clinical discussion rather than a formal standardized capacity-assessment instrument. Patients were informed that participation was entirely voluntary, that refusal or subsequent withdrawal would not affect their hospitalization, psychiatric treatment, or access to routine clinical care, and that study participation was separate from clinical treatment decisions. Patients were enrolled only when they were considered clinically able to provide informed consent.

The same psychiatrist, with 25 years of experience in psychiatry and training in PANSS administration and scoring provided by Signant Health, conducted the baseline psychopathological assessments using the Positive and Negative Syndrome Scale (PANSS) for all included participants. Neuropsychological assessments were performed independently by one clinical psychologist (Master's degree; 15 years of experience in clinical psychology and psychometric assessment), who administered and scored the Benton Visual Retention Test (BVRT) at baseline (T1) and at the seven-day follow-up (T2). The psychologist did not take part in the PANSS ratings and was not informed of the PANSS scores, but was aware of the participants' diagnoses. The same BVRT administration and scoring procedures were applied across both wards and assessment occasions.

Recruitment was conducted in parallel across the male and female wards, with a target of 100 participants with complete paired assessments per ward. This recruitment strategy resulted in equal representation of female and male participants in the final analytical sample. To prevent duplicate enrollment, hospital admission records and available patient identifiers were reviewed during eligibility screening to identify previous participation in the study. Individuals who had already been enrolled were not included again in subsequent admissions. Consequently, each participant contributed only one baseline and one seven-day follow-up BVRT assessment, and the final analytical sample comprised 200 distinct individuals.

Eligible participants were adults aged 18 years or older with a clinical diagnosis of schizophrenia or acute and transient psychotic disorders (ICD-10 F23), secondary or higher educational attainment, and written informed consent. Educational attainment was classified according to the highest completed level of formal education, based on patient declaration on admission. The requirement for at least secondary education was applied in order to understand the tasks. Although educational background may influence neuropsychological performance, completion of secondary education was not considered an intrinsic requirement for performing the BVRT copying task. Eligibility was assessed through clinical evaluation, patient and collateral history, and medical records. Study-specific exclusion criteria included borderline intellectual functioning, diagnosed neurological disorders or previous neurosurgery, organic mental disorders, substance-induced mental disorders and chronic pain. This exclusion was not based on poor performance on the BVRT or on a score threshold derived from the study assessments. These conditions were not inferred from low BVRT performance. Substance-use history was distinguished from a documented substance-related diagnosis; it did not by itself redefine an otherwise eligible psychotic disorder [43].

Diagnostic classification was based on the International Classification of Diseases, 10th Revision (ICD-10), using clinical information documented during routine psychiatric hospitalization. Diagnoses of schizophrenia and acute and transient psychotic disorders (F23) were established by the responsible treating psychiatrist through clinical evaluation,

including assessment of the presenting psychopathology, patient and collateral history, and available previous medical records. Diagnostic information was documented at admission and was available before enrollment and baseline study assessments, which were performed 48–72 h after admission. The psychiatrist responsible for study recruitment reviewed the available clinical documentation and verified diagnostic eligibility according to the predefined inclusion and exclusion criteria. The same diagnostic ascertainment and eligibility verification procedures were applied to participants in both the schizophrenia and F23 groups. The F23 group was defined according to ICD-10 criteria and was not restricted to first-episode psychosis.

Of 248 patients screened, 48 did not contribute paired assessments: 18 were discharged before T2; 12 reported significant drug use and had documented substance-related diagnoses; seven withdrew or refused the second assessment; seven had documented neurological problems with an associated organic component; and four had documented borderline intellectual functioning. These categories account for all 48 non-included patients. The available screening summary did not establish the total number who completed T1 before non-inclusion. The paired sample comprised 113 participants with schizophrenia and 87 with F23 disorders (Figure 1). Participant flow distinguishes clinical exclusions from noncompletion, consistent with STROBE reporting principles [45,46].

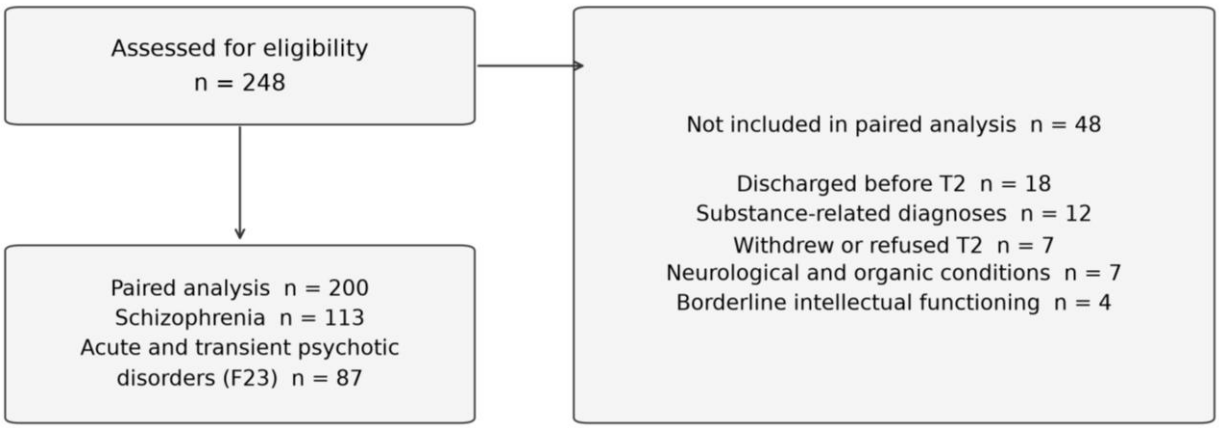


Figure 1. Participant flow from screening to the paired analysis. The five categories account for all 48 non-included patients. The screening summary does not establish the number who completed T1 before non-inclusion. The analysis includes 200 complete pairs without imputed T2 scores.

The initial neuropsychological assessment was performed 48–72 h after admission, following treatment initiation, when participants were considered sufficiently alert and capable of engaging in the assessment. Readiness for testing was determined by the psychiatrist, based on the ability to understand instructions, maintain adequate attention, and participate meaningfully in testing. Participants who were not sufficiently able to engage in testing had the assessment postponed and participated in the 2nd wave or were excluded from the study.

Testing was conducted in the psychiatrist's cabinet (both in T1 and T2). BVRT administration was standardized across participants and assessment occasions. All participants received the same task instructions and were assessed using BVRT Form C, Administration C (copying), with the same ten stimulus cards presented at both assessments. Each stimulus remained visible throughout the copying procedure, and participants were instructed to reproduce the figures as accurately as possible. The same administration and scoring procedures were applied at baseline (T1) and at the seven-day follow-up (T2).

2.2. Assessment and Scoring

Baseline assessments (T1) were conducted 48–72 h after admission and after treatment initiation, when participants were considered sufficiently alert and clinically able to complete the evaluation. Baseline psychopathological severity was assessed using the Positive and Negative Syndrome Scale (PANSS) by the same psychiatrist for all participants. The Benton Visual Retention Test (BVRT), Form C, Administration C (copying), was administered individually by the same clinical psychologist at baseline (T1) and at the seven-day follow-up (T2) [38,39]. Participants were instructed to reproduce the figures presented on each card as accurately as possible, while the stimulus card remained visible throughout the copying procedure. The same ten stimulus cards were presented at both assessments.

Both assessments were conducted under untimed conditions, allowing participants to complete each drawing at their own pace, without a predefined time limit. No corrective feedback regarding drawing accuracy or specific errors was provided following the baseline assessment (T1), thereby avoiding additional examiner-directed learning between assessments.

Both T1 and T2 assessments were conducted in the psychiatrist's consulting room, rather than at the bedside or in other ward spaces. The same psychologist administered the test using identical instructions, stimulus materials, administration procedures, and scoring criteria at both assessment occasions. These procedures were intended to maintain consistency across assessments and diagnostic groups, although potential practice effects resulting from repeated exposure to identical stimuli could not be eliminated.

No examiner-directed assistance, corrective prompting, or guidance regarding the accuracy or spatial organization of the drawings was provided during either assessment. Participants were permitted to make spontaneous corrections to their own drawings.

Each whole card generated one response, scored as correct or incorrect, irrespective of its number of component figures. The primary outcome was the study-recorded BVRT copying correct-card count, derived from the binary classification of each of the ten stimulus cards as correct or incorrect (range: 0–10). Each card contributed one point when classified as correct, irrespective of the number of component figures. The same examiner and scoring approach were used at both assessment occasions.

Qualitative drawing observations, including omissions, distortions, perseverations, rotations, misplacements, and size errors, were documented independently of the binary correct-card classification. These observations were coded from the completed drawings. A qualitative observation did not automatically result in a card being classified as incorrect.

However, the original detailed tolerance criteria used to distinguish acceptable drawing variations from errors requiring an incorrect-card classification could not be recovered from the available scoring documentation. In particular, the records did not establish the precise criteria used to classify cards as correct when qualitative observations were present. Therefore, the primary outcome is described as a study-recorded BVRT copying score rather than a fully verified manual-standardized BVRT score. This distinction limits direct comparability with studies employing explicitly documented standardized BVRT correctness criteria.

Qualitative observations were recorded separately under omissions, distortions, perseverations, rotations, misplacements and size errors. Under the study recording procedure, these observations did not determine or modify the binary correct/incorrect classification and could accompany a response classified as correct. Several observations could occur within one incorrect response; for example, an omission and a rotation would contribute one incorrect response and two qualitative observations. Thus, the qualitative counts need not equal the number of incorrect responses or sum to ten. They were not

used to reconstruct the correct-card score and were not interpreted as a standardized BVRT Number Error Score.

PANSS positive (P), negative (N) and general psychopathology (G) subscales were analyzed as recorded. Total PANSS was P + N + G; the composite score was P minus N and did not represent overall severity [47–49]. Covariates were age, recorded sex, education and diagnosis. Auditory, visual and mixed hallucination codes were mutually exclusive; delusions were coded separately. Treatment was represented by two broad categories recorded in the clinical file, following the traditional clinical grouping of antipsychotics by predominant profile: "incisive" (antipsychotics prescribed mainly for their antipsychotic action—risperidone, paliperidone, aripiprazole, amisulpride) and "sedative" (antipsychotics prescribed mainly for their sedative action—olanzapine, quetiapine, clozapine). Participants receiving more than one drug were assigned to a single category according to the principal antipsychotic recorded in the treatment plan at T1; adjunctive benzodiazepines, hypnotics and anticholinergic agents were not considered. The classification reflects treatment at T1; drug-specific doses, exposure histories and changes between T1 and T2 were not available. Hallucination and delusion codes were obtained from the admission psychiatric examination documented in the medical record and by the clinical evaluation, were verified by the study psychiatrist at enrollment, describe the current episode up to T1, and were not derived from PANSS item scores. Hallucination codes were mutually exclusive: auditory only, visual only, mixed (two or more sensory modalities), or none documented. Delusions were coded separately as present or absent. These variables were reserved for exploratory analyses.

Data Sources, Recording Procedures, and Derivation of Analytical Variables

Data were obtained from study-related neuropsychological and psychopathological assessments and from clinical information available during participant evaluation. Following informed consent, the baseline assessment (T1) was conducted 48–72 h after hospital admission, and the Benton Visual Retention Test (BVRT) was repeated seven days later (T2). Both BVRT assessments were administered by the clinical psychologist using Form C, Administration C, with identical stimulus cards and administration procedures.

The original source records for BVRT performance were the participants' completed drawing sheets and the psychologist's scoring sheets. Each card was classified as correctly or incorrectly copied from the original drawings.

Six qualitative drawing-observation categories—omissions, distortions, perseverations, rotations, misplacements, and size errors—were documented separately from the binary correct-card classifications. These observations were coded retrospectively from the retained original drawings by the clinical psychologist. The qualitative observations were not used to reconstruct or modify the correct-card scores and were not treated as a standardized BVRT Number Error Score.

Six categories of qualitative drawing observations were evaluated separately from the binary whole-card correct/incorrect classification: omissions, distortions, perseverations, rotations, misplacements, and size errors. Omissions referred to absent stimulus elements or figure components; distortions to alterations in figure configuration or contour; perseverations to unintended repetitions or continuations of drawing elements; rotations to deviations in figure orientation; misplacements to incorrect spatial positioning of figures or components; and size errors to discrepancies in relative size or proportions.

The counting unit was an individual, distinguishable observation rather than an affected card. Each separate occurrence was counted within its corresponding category, and the same category could be recorded more than once on a card when distinct occurrences were identified. A single occurrence was not counted repeatedly within the same

category. Different categories could be assigned to the same card when distinct abnormalities were present.

Because qualitative observations were coded retrospectively from the completed drawings, intermediate drawing behaviour and spontaneous corrections made during copying were not available for coding. Whole-card correctness was evaluated separately from the qualitative observations, which were not used to calculate or modify the binary correct-card score. The qualitative counts were therefore considered study-specific descriptive observations rather than a standardized BVRT Number Error Score.

Baseline PANSS assessments were performed by the study psychiatrist using the Romanian version of the 30-item PANSS, validated in the Romanian population, with ratings referring to the past week, the standard PANSS rating period [47]. PANSS positive, negative, and general psychopathology subscale scores were obtained from the original PANSS rating forms completed by the psychiatrist at T1. Demographic and diagnostic information was obtained through clinical evaluation, patient and collateral history, and available medical records. The sources, timing and coding rules for the treatment, hallucination and delusion variables are described in Section 2.2.

Following data collection, the available observations were assembled into the analytical dataset and underwent consistency checks. Derived variables were the number of correctly copied cards at each assessment (from the card-level classifications), the within-participant BVRT change score (T2 minus T1), the PANSS total (P + N + G) and the PANSS composite score (P minus N).

Subsequent statistical analyses included paired comparisons, correlation and regression analyses, and additional exploratory and post hoc investigations. In particular, the extended multivariable analyses and longitudinal Poisson generalized estimating equations analyses of qualitative observations were conducted after the original assessments and were not prospectively prespecified. These analytical procedures did not modify the originally recorded BVRT scores.

2.3. Statistical Analysis

Participants were considered the independent sampling units. Continuous variables were summarized using means and standard deviations (SDs), categorical variables were presented as absolute frequencies and percentages. The primary outcome was the within-participant change in the number of correctly copied Benton Visual Retention Test (BVRT) cards between baseline (T1) and the seven-day follow-up assessment (T2), calculated as T2 minus T1. Mean changes were estimated using Student's t-distribution to construct 95% confidence intervals (CIs), and statistical significance was assessed using a two-sided paired Student's t-test [50,51]. The standardized effect size was expressed as Cohen's dz, calculated by dividing the mean within-participant change by the standard deviation of the paired differences [52–55]. The continued relevance of paired comparisons and standardized effect-size estimation in repeated-measures research has been emphasized in recent methodological investigations [51,56].

To evaluate the robustness of the primary findings to the discrete and non-normally distributed difference scores, a percentile bootstrap procedure with 20,000 participant-level resamples and the Wilcoxon signed-rank test were additionally performed [56–58]. For the Wilcoxon analysis, zero differences were omitted before ranking, the asymptotic variance was adjusted for tied absolute differences, and no continuity correction was applied. The treatment of zero differences and tied ranks followed established methodological considerations regarding signed-rank procedures [59,60]. These complementary analyses were used to assess the stability of the observed change under alternative statistical assumptions.

Associations between baseline psychopathological symptom severity and concurrent visuoconstructive copying performance were initially assessed using Spearman's rank correlation coefficient [61,62]. The association between the Positive and Negative Syndrome Scale (PANSS) negative subscale (PANSS N) and baseline BVRT performance was evaluated using a 20,000-resample bootstrap confidence interval and a two-sided permutation test based on 20,000 permutations. Permutation p values were calculated as (b + 1)/20,001, where b represented the number of permuted statistics at least as extreme as the observed statistic. Pearson's product–moment correlation coefficient was additionally calculated as a sensitivity analysis [62,63]. Both correlation approaches remain relevant in contemporary investigations examining associations between psychopathological symptom dimensions and visual processing performance in psychotic disorders [62].

Multivariable associations with seven-day BVRT copying performance were examined using ordinary least squares (OLS) regression. The primary longitudinal regression model estimated BVRT performance at T2 from baseline PANSS N scores, baseline BVRT performance, age, sex, educational level, and psychiatric diagnosis, with all predictors entered simultaneously. Female sex, secondary education, and acute and transient psychotic disorders (ICD-10 F23) were the reference categories. Heteroscedasticity-consistent HC3 covariance estimates were used to obtain robust standard errors, confidence intervals, and p values, following established procedures for regression inference under potential heteroscedasticity [64,65]. Statistical inference was based on the residual t-distribution with 193 degrees of freedom. The importance of examining regression assumptions and interpreting adjusted associations appropriately has also been emphasized in recent health research [66].

The longitudinal regression model was specified as follows:

$$\text{BVRT T2} = \beta_0 + \beta_1(\text{PANSS N at T1}) + \beta_2(\text{BVRT T1}) + \beta_3(\text{age}) + \beta_4(\text{male sex}) + \beta_5(\text{higher education}) + \beta_6(\text{schizophrenia}) + \varepsilon.$$

Each participant contributed one T2 outcome, with baseline BVRT performance included as a covariate. PANSS total and composite scores were not entered alongside the PANSS subscales because these variables were derived from the same underlying symptom scores. Adjustment for baseline performance was intended to account for differences in initial copying accuracy; however, given the observational design, the resulting regression coefficients were interpreted as associations rather than causal effects.

Regression model assumptions were investigated through graphical examination of residual distributions and residual-versus-fitted plots, assessment of heteroscedasticity using the Breusch–Pagan test [66,67], and evaluation of multicollinearity using variance inflation factors [66,68]. Because the BVRT correct-card score is bounded between zero and ten, a fractional-logit model was additionally fitted as a bounded-outcome sensitivity analysis. The dependent variable was expressed as the proportion of correctly copied cards (BVRT T2/10), and the same covariates as in the primary longitudinal model were included. Heteroscedasticity-consistent HC0 robust covariance estimates were used for inference. This approach was based on fractional-response regression methodology developed by Papke and Wooldridge (1996) [69] and is consistent with more recent methodological recommendations concerning the analysis of bounded-count health outcomes.

Additional post hoc analyses were conducted to characterize baseline performance, examine potential nonlinear associations, and assess the robustness of the main findings. These analyses included descriptive comparisons of BVRT changes across baseline performance strata, an adjusted model examining the concurrent association between PANSS N and baseline BVRT performance, an extension of the longitudinal regression model incorporating a quadratic baseline BVRT term, and sensitivity analyses involving historical data annotations. The post hoc concurrent regression model included BVRT T1 as the

dependent variable and PANSS N, age, sex, educational level, and diagnosis as independent variables, with HC3 robust inference based on 194 residual degrees of freedom. The quadratic extension of the longitudinal model incorporated the squared, mean-centered baseline BVRT score to investigate potential curvature in the relationship between initial and follow-up performance. These analyses were not used for stepwise variable selection or the development of a clinical prediction model. No BVRT scores were modified during these procedures. Arithmetic consistency checks were performed for PANSS total and composite scores and BVRT scores, and complete paired records were analyzed without imputation.

Exploratory analyses investigated associations involving the PANSS positive (PANSS P) and general psychopathology (PANSS G) subscales, diagnostic-group contrasts, changes in the six qualitative drawing-observation categories, and treatment- or symptom-related indicators. To reduce the risk of false-positive findings arising from multiple comparisons, the Benjamini–Hochberg false-discovery-rate correction was applied within separate families of two, two, six, and six statistical tests, respectively [62,70]. The final six-test family comprised five separately adjusted treatment or symptom exposure tests and one joint hallucination test. Wilcoxon analyses of qualitative observation counts constituted a separate family of six tests. Multiplicity-adjusted p values were denoted by q, whereas confidence intervals were reported as individual, unadjusted intervals. The primary paired comparison and the two secondary analyses involving PANSS N were evaluated at a nominal two-sided significance threshold of $\alpha = 0.05$. Although this analytical hierarchy reflected the structure of the present investigation, prospective prespecification of the hierarchy was not documented.

Further post hoc analyses were undertaken to characterize the relationship between psychopathological symptom dimensions and baseline copying performance. An extended OLS regression model examined the concurrent association between PANSS N and BVRT T1 while simultaneously adjusting for PANSS P, PANSS G, age, sex, educational level, and diagnosis. Heteroscedasticity-consistent HC3 standard errors and the residual t distribution (192 degrees of freedom) were used for statistical inference, as for the other linear models. This model was intended to determine whether the concurrent association between conventional PANSS negative-subscale severity and copying performance remained evident after accounting for other measured psychopathological dimensions and demographic and diagnostic characteristics. A further post hoc analysis, added during revision, replaced or supplemented PANSS N with the PANSS anergia (lack of energy) cluster score [34], which does not include the cognitive items N5 (difficulty in abstract thinking) and N7 (stereotyped thinking), using the same covariates as Model A, HC3 standard errors and the residual t distribution.

The longitudinal evolution of the six separately coded qualitative drawing-observation categories—omissions, distortions, perseverations, rotations, misplacements, and size errors—was additionally examined using a Poisson generalized estimating equations (GEE) model. This approach follows the framework introduced by Liang and Zeger (1986) [71] for analyzing correlated longitudinal responses and remains relevant in contemporary clinical research involving repeated or clustered observations [72,73]. The model used a log link and a Poisson variance function, with data arranged as one record per participant, assessment and category (12 records per participant; 2400 records from 200 participants; no missing values). The mean model included assessment time, qualitative observation category and their interaction (a saturated mean model with 12 parameters), with T1 and omissions as reference levels. Repeated observations were clustered at the participant level, with an exchangeable working correlation structure and robust sandwich standard errors. A joint Wald test with five degrees of freedom evaluated the overall time-by-category interaction, that is, whether relative changes in qualitative observation counts

differed across categories [73,74]. Category-specific T2/T1 time ratios were estimated with robust 95% confidence intervals. Because the mean model was saturated and the data were complete and balanced, the fitted means equal the observed category-by-visit means, and the robust variance does not depend on the working correlation structure; an alternative working correlation would therefore yield identical estimates and tests. The finite-sample stability of the interaction test was instead assessed with a bias-reduced (Mancl–DeRouen) sandwich variance estimator [75,76] and with a Wald test based on a participant-level cluster-bootstrap covariance of the interaction coefficients (20,000 resamples).

Additionally, all 15 pairwise contrasts between within-participant changes in the six qualitative observation categories were assessed using paired t-tests, with Benjamini–Hochberg adjustment applied across the 15 comparisons. These supplementary longitudinal analyses were exploratory, were not prospectively prespecified, and were interpreted separately from the primary and secondary outcomes.

Diagnostic-group comparisons were conducted using Welch's unequal-variance t-test to compare mean BVRT change scores between participants with schizophrenia and those with acute and transient psychotic disorders [77,78]. The diagnostic coefficient from the adjusted T2 regression model was also examined. These comparisons addressed distinct estimands: the unadjusted between-group difference in mean change and the adjusted difference in follow-up performance conditional on baseline BVRT scores and other covariates. Changes in qualitative observation counts were examined using paired t-tests, with separate Wilcoxon signed-rank tests performed as sensitivity analyses.

Exploratory associations involving treatment and clinical symptom indicators were investigated by adding each of the five indicators separately to the base longitudinal regression model. An additional model simultaneously included auditory, visual, and mixed hallucination categories, with the absence of coded hallucinations serving as the reference category. The overall association of hallucination categories with follow-up BVRT performance was evaluated using a robust global F-test. In models examining hallucination indicators separately, the absence of one specific hallucination category could include participants assigned to another hallucination category. Accordingly, these exploratory comparisons were interpreted cautiously and were not considered evidence of causal treatment effects or symptom-related changes in cognitive functioning.

2.4. Sample Size and Sensitivity

No prospective sample size calculation was documented. A post hoc sensitivity calculation used nominal two-sided alpha = 0.05 and the noncentral t distribution to determine minimum detectable effects at 80% and 90% power. The paired calculation used n = 200 and 199 degrees of freedom; dz was the mean difference divided by the SD of differences. Conversion to BVRT points assumed the observed difference SD of 1.5556. A separate equal-variance approximation used diagnostic-group sizes of 113 and 87, a pooled change SD of 1.5492 and 198 degrees of freedom. The reported diagnostic contrast itself used Welch inference. The observed mean effect was not inserted as a target to report "observed power" [79,80]. These calculations are not power estimates for the adjusted PANSS coefficient or multiplicity-adjusted exploratory tests and do not establish clinical importance.

2.5. Software and Data Checks

Analyses used Python 3.12.14 [81], pandas 2.2.3 [82,83], NumPy 2.5.4 [84,85], SciPy 1.18.1 [86,87], statsmodels 0.15.0 [88] and Matplotlib 3.11.2 [89,90]. Random-number generation for bootstrap resampling and permutation analyses was performed using NumPy's default_rng function, initialized with a fixed seed of 20260913 to facilitate computational reproducibility [84,91]. The generator used the default PCG64 bit generator. A

total of 20,000 participant-level bootstrap resamples and 20,000 permutations per PANSS subscale were generated according to the procedures described above. For computational reproducibility, the initial array of 20,000 × 200 participant indices generated using NumPy's default_rng function was used for the bootstrap estimation of the mean BVRT change and subsequently reused for all correlation bootstrap confidence intervals. Spearman ranks were recalculated within each bootstrap resample. Permutation analyses were subsequently performed using the same continuing random-number stream, in the order PANSS N, PANSS P, and PANSS G, with 20,000 permutations per subscale. Two-sided permutation *p* values were calculated as (b + 1)/20,001, where b represents the number of permutations yielding a test statistic at least as extreme as the observed statistic. All numerical estimates were rounded only after completion of the statistical calculations.

Data checks confirmed unique study identifiers, complete main-model variables, correct + incorrect = 10 at both visits, delta = T2 minus T1, PANSS total = P + N + G, and PANSS composite = P minus N. The locked analysis dataset (version of 9.10.2026) contained audit flags that are documented in the following data-audit note and defined in the data dictionary (Supplementary File S1). Qualitative observation counts were coded separately from binary correctness and were not grounds for excluding a participant or altering a BVRT score.

Data-audit note: during data checking, all PANSS subscale and total scores were compared with the original PANSS rating forms. In one record (BVRT183), the recorded PANSS total (119) did not equal the sum of the subscale scores (P 23 + N 29 + G 55 = 107). The subscale scores matched the rating form, and the discrepancy was attributed to an arithmetic error in the total, which was corrected to 107 by the psychiatrist. The correction affected only the derived PANSS total; the P, N and G subscales and the P minus N composite were unchanged, so no analysis based on the subscales was affected. A further legacy marker identified 36 records in which, at one or both visits, fewer qualitative observations than incorrect cards were recorded; because the two measures were coded separately, this marker was retained for provenance only and was not treated as a data error. All flagged records were retained in the primary analysis; sensitivity analyses omitting the corrected and annotated records are reported in Section 3.6.

Additional analyses performed during revision (Supplementary Table S1 Panel C and Supplementary Table S2) used Python 3.13.3 with the same NumPy, SciPy and statsmodels versions and pandas 3.1.1; the bootstrap used NumPy default_rng with seed 20260913 and 20,000 participant-level resamples.

3. Results

3.1. Participant Flow and Baseline Characteristics

The paired analysis included 200/248 screened patients (80.6%); 48 (19.4%) were not included for the reasons shown in Figure 1 (created in Word by the authors—participant-flow diagram constructed to illustrate the numbers of patients screened, excluded, and included in the final analysis, together with the distribution of participants by diagnostic category). All analyzed participants had complete data for the main model. The sample comprised 100 women and 100 men, with a mean age of 39.52 years (SD 10.82). The schizophrenia group was older and had higher mean PANSS N and total scores than the F23 group (Table 1). These baseline differences are relevant to interpreting unadjusted diagnostic comparisons.

Table 1. Participant characteristics and recorded clinical measures by diagnosis.

Characteristic	Total (n = 200)	Schizophrenia (n = 113)	F23 (n = 87)
Age in years	39.52 (10.82)	42.54 (9.65)	35.59 (11.05)
Female sex	100 (50.0%)	54 (47.8%)	46 (52.9%)
Male sex	100 (50.0%)	59 (52.2%)	41 (47.1%)
Secondary education	139 (69.5%)	82 (72.6%)	57 (65.5%)
Higher education	61 (30.5%)	31 (27.4%)	30 (34.5%)
PANSS P at T1	26.11 (4.49)	26.80 (4.93)	25.21 (3.69)
PANSS N at T1	27.51 (4.54)	29.19 (4.00)	25.32 (4.29)
PANSS G at T1	57.99 (6.38)	59.20 (6.46)	56.41 (5.96)
PANSS total at T1	111.60 (12.92)	115.19 (13.27)	106.94 (10.86)
PANSS composite P minus N	−1.40 (4.89)	−2.39 (4.91)	−0.11 (4.58)
Correctly copied cards at T1	6.52 (2.90)	6.19 (2.91)	6.95 (2.84)
Maximum score at T1	25 (12.5%)	10 (8.8%)	15 (17.2%)
Incisive treatment category	138 (69.0%)	78 (69.0%)	60 (69.0%)
Sedative treatment category	62 (31.0%)	35 (31.0%)	27 (31.0%)
Auditory hallucinations	69 (34.5%)	47 (41.6%)	22 (25.3%)
Visual hallucinations	15 (7.5%)	8 (7.1%)	7 (8.0%)
Mixed hallucinations	64 (32.0%)	34 (30.1%)	30 (34.5%)
No coded hallucinations	52 (26.0%)	24 (21.2%)	28 (32.2%)
Delusions	144 (72.0%)	88 (77.9%)	56 (64.4%)

Note: Values are mean (SD) or n (% within column). F23, acute and transient psychotic disorders; PANSS, Positive and Negative Syndrome Scale. PANSS total = P + N + G; composite = P minus N. Treatment categories reproduce broad source classifications and are mutually exclusive. Hallucination categories are mutually exclusive; delusions were coded separately. Group differences are descriptive and were not screened by baseline significance tests.

3.2. Change in Copying Performance over Seven Days

Mean correct-card scores increased from 6.52 (SD 2.90) to 7.36 (SD 2.84), a mean change of 0.835 points (95% CI 0.618–1.052; t(199) = 7.591; $p < 0.001$; dz = 0.537). The bootstrap CI was 0.625–1.050, and the Wilcoxon test supported a shift in scores (W = 1760; $p < 0.001$). Scores increased in 120 participants (60.0%), were unchanged in 58 (29.0%) and decreased in 22 (11.0%). Table 2 and Figure 2 (generated using Python (Matplotlib 3.11.2). The left panel was constructed as a two-dimensional frequency heatmap of paired baseline and follow-up scores, while the right panel was generated as a frequency distribution of within-participant score differences (T2 minus T1)) describe the paired results.

Table 2. Paired BVRT copying outcomes.

Measure	Estimate
Correctly copied cards at T1, mean (SD)	6.52 (2.90)
Correctly copied cards at T2, mean (SD)	7.36 (2.84)
Incorrectly copied cards at T1, mean (SD)	3.48 (2.90)
Incorrectly copied cards at T2, mean (SD)	2.65 (2.84)
Mean change (SD)	0.835 (1.556)
95% CI for mean change	0.618 to 1.052
Percentile bootstrap 95% CI	0.625 to 1.050
Paired t statistic, df and p	7.591; 199; <0.001
Wilcoxon W and p	1760; <0.001
Cohen's dz	0.537
Median change (Q1, Q3)	1.00 (0.00, 1.25)

Higher/unchanged/lower at T2	120 (60.0%)/58 (29.0%)/22 (11.0%)
Maximum score at T1/T2	25 (12.5%)/57 (28.5%)

Note: n = 200. Change is T2 minus T1 in correctly copied whole cards. Correct plus incorrect responses equals ten at each assessment; rounded displayed means may differ from ten by 0.01. Wilcoxon concerns ranked differences rather than the mean. Cohen's dz uses the SD of within-participant differences and is not a controlled treatment effect.

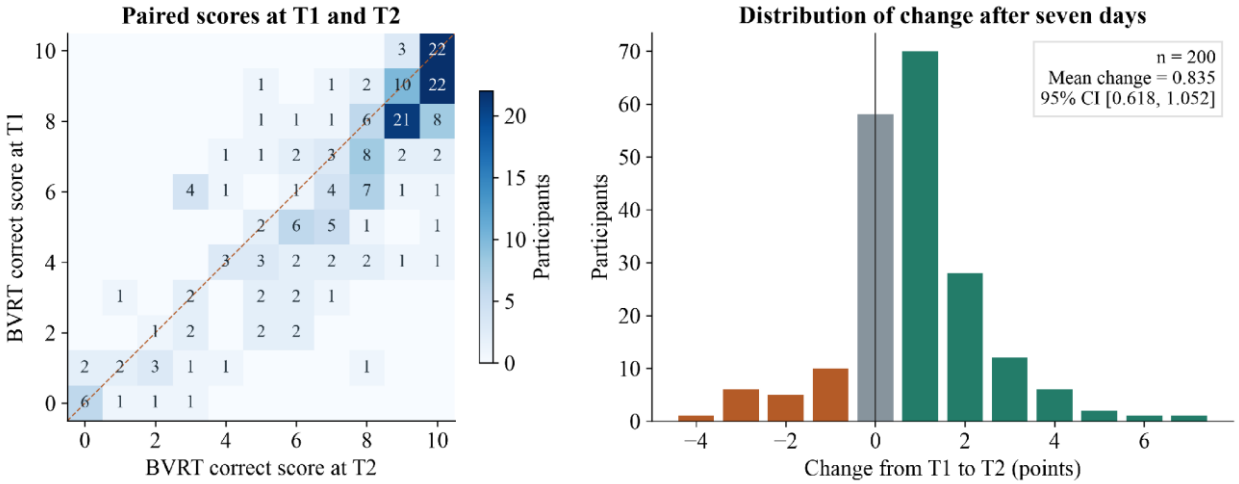


Figure 2. BVRT Form C, Administration C, seven days apart. (**Left**): numbers of participants at each paired score; the diagonal indicates no change. (**Right**): distribution of individual changes. Each score counts correctly copied whole cards from zero to ten. Positive change indicates a higher T2 score.

3.3. Baseline Performance and Ceiling

Maximum scores increased from 25 (12.5%) at T1 to 57 (28.5%) at T2. Among the 25 participants initially scoring ten, 22 remained at ten and three declined. Post hoc baseline strata showed larger mean gains at lower initial scores (Table 3); this descriptive pattern is constrained by the score ceiling and cannot distinguish regression to the mean from other influences.

Table 3. Post hoc change by baseline correct-card score.

T1 Score	n	Mean Change	95% CI	Higher	Same	Lower
0–4	48	1.646	1.092 to 2.200	31	14	3
5–7	53	0.868	0.375 to 1.361	38	6	9
8–9	74	0.608	0.360 to 0.856	51	16	7
10	25	−0.120	−0.257 to 0.017	0	22	3

Note: Change is T2 minus T1. Strata are descriptive and were formed after inspecting the score distribution; no between-stratum hypothesis test was used. The ceiling prevents positive changes from a baseline of ten. Baseline–change patterns may reflect headroom, measurement variation, regression to the mean and mathematical coupling. They do not identify treatment responsiveness.

For all participants below ten at T1 (n = 175), mean change was 0.971 points (95% CI 0.731–1.212); 120 increased, 36 were unchanged and 19 decreased. This post hoc subgroup estimate changes the population being described and is not a replacement for the full-sample estimate. At T1, zero scores occurred in nine participants; at T2, in eight. The change range was −4 to 7, and Shapiro–Wilk W for differences was 0.916 ($p < 0.001$).

3.4. PANSS and Concurrent Copying Performance

PANSS N was inversely associated with BVRT T1 (Spearman rho = −0.313; bootstrap 95% CI −0.440 to −0.176; permutation $p < 0.001$). Pearson r was −0.341 (95% CI −0.458 to −0.212). A post hoc model adjusting the concurrent association for age, sex, education and diagnosis yielded B = −0.197 BVRT points per PANSS N point (95% CI −0.317 to −0.077; $p = 0.0014$). PANSS P and G exploratory correlations were each approximately −0.222 (Tables 4 and 5). Figure 3 displays the concurrent PANSS N association (generated using Python (Matplotlib 3.11.2) as a frequency-weighted scatter plot, with marker sizes proportional to the number of participants sharing each score combination).

Table 4. PANSS subscales and BVRT T1.

Subscale	Spearman rho	Bootstrap 95% CI	Permutation *p*	BH q
PANSS N	−0.313	−0.440 to −0.176	<0.001	Not applied
PANSS P	−0.222	−0.362 to −0.077	0.0023	0.0023
PANSS G	−0.222	−0.354 to −0.080	0.0020	0.0023

Note: n = 200. Bootstrap and permutation analyses each used 20,000 iterations. The minimum attainable permutation *p* was 1/20,001. PANSS N addressed a secondary objective; BH adjustment applies to P and G only. Pearson sensitivity for PANSS N gave r = −0.341 (Fisher 95% CI −0.458 to −0.212).

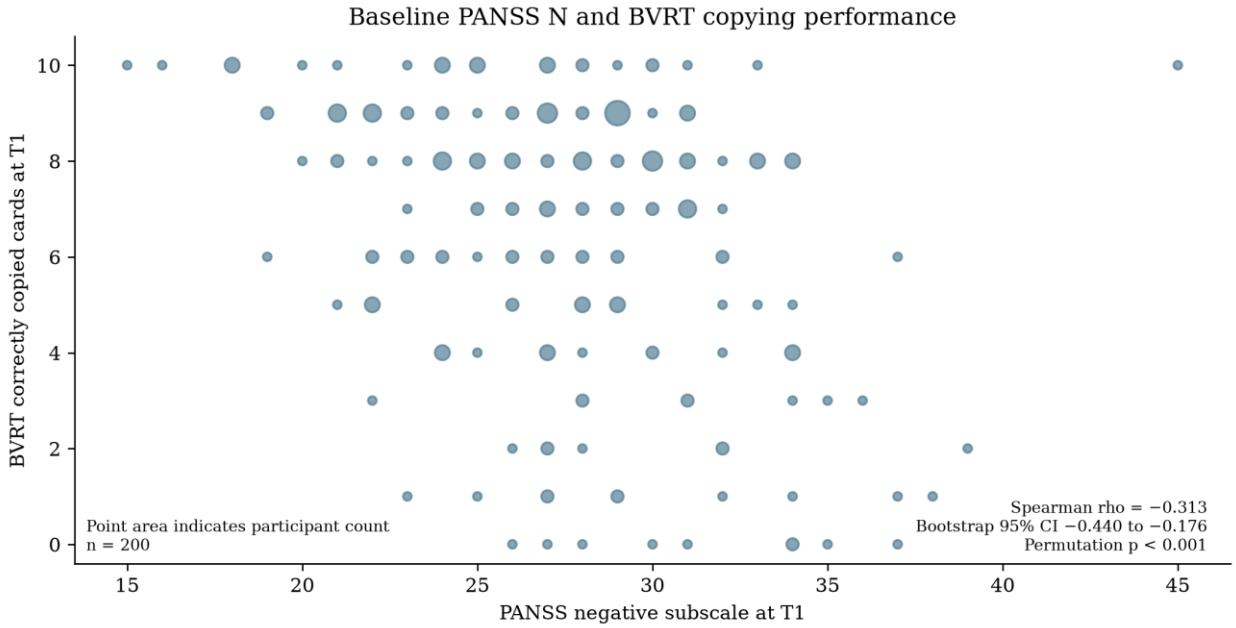


Figure 3. Baseline PANSS N and BVRT T1. Point area represents the number of participants with the same pair of scores. The Spearman estimate and bootstrap interval describe a concurrent association at T1.

Both post hoc concurrent models appear in Table 5. Model A adjusts the PANSS N coefficient for age, sex, education and diagnosis; Model B (Section 3.5) additionally adjusts for PANSS P and PANSS G. It addresses a different question from the model of retest performance conditional on baseline BVRT.

Table 5. Post hoc adjusted concurrent associations with baseline BVRT.

Predictor or Contrast	Model A: B (HC3 95% CI)	*p*	Model B: B (HC3 95% CI)	*p*
Intercept	10.787 (7.735 to 13.838)	<0.001	11.598 (6.569 to 16.626)	<0.001
PANSS N at T1 per point	−0.197 (−0.317 to −0.077)	0.0014	−0.170 (−0.301 to −0.039)	0.011

PANSS P at T1 per point	-	–	−0.095 (−0.211 to 0.021)	0.109
PANSS G at T1 per point	-	–	0.013 (−0.089 to 0.115)	0.808
Age per year	0.015 (−0.028 to 0.059)	0.490	0.020 (−0.023 to 0.063)	0.361
Male versus female	0.493 (−0.287 to 1.273)	0.214	0.502 (−0.277 to 1.281)	0.205
Higher versus secondary education	1.084 (0.279 to 1.888)	0.009	1.142 (0.334 to 1.949)	0.006
Schizophrenia versus F23	−0.063 (−0.995 to 0.870)	0.894	−0.080 (−1.020 to 0.860)	0.867
Residual df	194		192	
R^2 (adjusted R^2)	0.150 (0.128)		0.165 (0.135)	
Breusch–Pagan *p*	0.005		<0.001	
Maximum VIF	1.391		1.996	

Note: n = 200. The outcome is BVRT T1 (correctly copied cards). Model A adjusts for age, sex, education and diagnosis; Model B additionally adjusts for PANSS P and PANSS G. Both models were added post hoc to examine whether the concurrent association persisted after adjustment; neither is the main longitudinal model. Reference categories were female sex, secondary education and F23. Inference uses HC3 covariance and the residual t distribution; Breusch-Pagan tests indicated heteroscedasticity. Pearson correlations between PANSS subscales were 0.414 (N–P), 0.589 (N–G) and 0.597 (P–G). VIF, variance inflation factor. Intervals and *p* values are nominal and are not adjusted for multiplicity.

3.5. Extended Multivariable Analysis of Baseline Copying Performance

In an additional post hoc model including all three PANSS subscales simultaneously, the inverse association between baseline PANSS N and BVRT copying performance remained evident after adjustment for PANSS P, PANSS G, age, sex, education, and diagnosis (Model B in Table 5: B = −0.170 BVRT points per PANSS N point; HC3 95% CI −0.301 to −0.039; $p = 0.011$). In the same model, PANSS P (B = −0.095; 95% CI −0.211 to 0.021; $p = 0.109$) and PANSS G (B = 0.013; 95% CI −0.089 to 0.115; $p = 0.808$) did not show clear additional associations with baseline copying performance. Model B explained 16.5% of the variance in baseline copying scores (adjusted $R^2 = 0.135$), compared with 15.0% for Model A; the maximum variance inflation factor was 1.996 (PANSS G), indicating no problematic multicollinearity despite correlations of 0.41–0.60 between the PANSS subscales. These findings suggest that the relationship between conventional PANSS negative-subscale severity and concurrent visuoconstructive performance was not fully accounted for by the other measured PANSS dimensions or demographic and diagnostic covariates. However, the association remained cross-sectional and did not establish a causal relationship or predict individual improvement.

The anergia cluster, which does not include N5 or N7, was also inversely associated with BVRT T1 (Spearman $\varrho = -0.250$; $p < 0.001$; adjusted B = −0.245 per point; 95% CI −0.445 to −0.044; $p = 0.017$). When PANSS N and anergia were entered together, PANSS N remained associated with BVRT T1 (B = −0.188; 95% CI −0.331 to −0.046; $p = 0.010$), whereas anergia did not (B = −0.026; 95% CI −0.251 to 0.200; $p = 0.823$); the two scores were correlated (Pearson r = 0.70).

3.6. Adjusted Association with Retest Performance

After adjustment for baseline BVRT and the other covariates, the PANSS N coefficient for T2 was −0.031 points (95% CI −0.087 to 0.026; $p = 0.286$; Table 6). For a ten-point PANSS N difference, the interval spans approximately 0.87 fewer to 0.26 more correctly copied cards. Model R^2 was 0.735 and adjusted R^2 was 0.726. Figure 4 displays the adjusted coefficients and their individual confidence intervals (generated using Python (Matplotlib 3.11.2) as a coefficient-and-confidence-interval plot based on the ordinary least squares regression estimates and HC3 robust covariance).

Table 6. Adjusted linear model of BVRT T2.

Predictor or Contrast	B	HC3 SE	95% CI	*p*
Intercept	2.797	0.912	0.998 to 4.595	0.002
PANSS N at T1 per point	−0.031	0.029	−0.087 to 0.026	0.286
BVRT T1 per point	0.823	0.042	0.740 to 0.906	<0.001
Age per year	−0.008	0.011	−0.029 to 0.014	0.492
Male versus female	0.206	0.236	−0.260 to 0.672	0.384
Higher versus secondary education	0.039	0.226	−0.408 to 0.485	0.865
Schizophrenia versus F23	0.385	0.222	−0.053 to 0.823	0.084

Note: n = 200; residual df = 193. R^2 = 0.735; adjusted R^2 = 0.726. B denotes unstandardised BVRT points. All predictors were entered simultaneously. Reference categories were female sex, secondary education and F23. Inference uses HC3 covariance estimates and the residual t distribution. Individual intervals are not adjusted for multiplicity; the exploratory diagnosis coefficient has BH q = 0.102 when considered with the unadjusted change contrast.

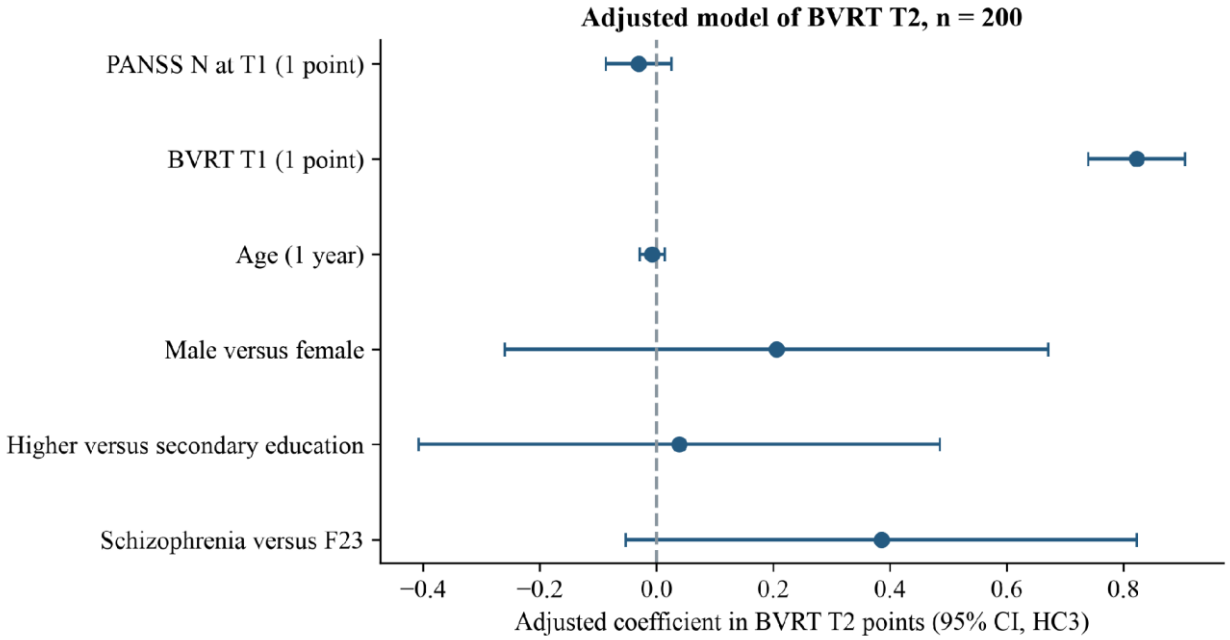


Figure 4. Adjusted associations with BVRT T2. Points show unstandardised coefficients and bars show individual HC3 95% confidence intervals from Table 6. The intercept is omitted. F23 denotes acute and transient psychotic disorders.

Model Robustness and Historical Data Flags

For the main linear T2 model, the Breusch–Pagan LM test gave p = 0.00536, and the maximum variance inflation factor was 1.503. Eighteen fitted outcomes exceeded ten and none was below zero. Robust covariance addresses standard-error estimation under heteroscedasticity but does not constrain fitted scores or establish the correct functional form. Figure 5 presents residual plots (generated using Python (Matplotlib 3.11.2) from the fitted ordinary least squares regression model. Residual-versus-fitted and normal quantile–quantile plots were constructed as graphical diagnostic procedures to complement the regression assumption checks.

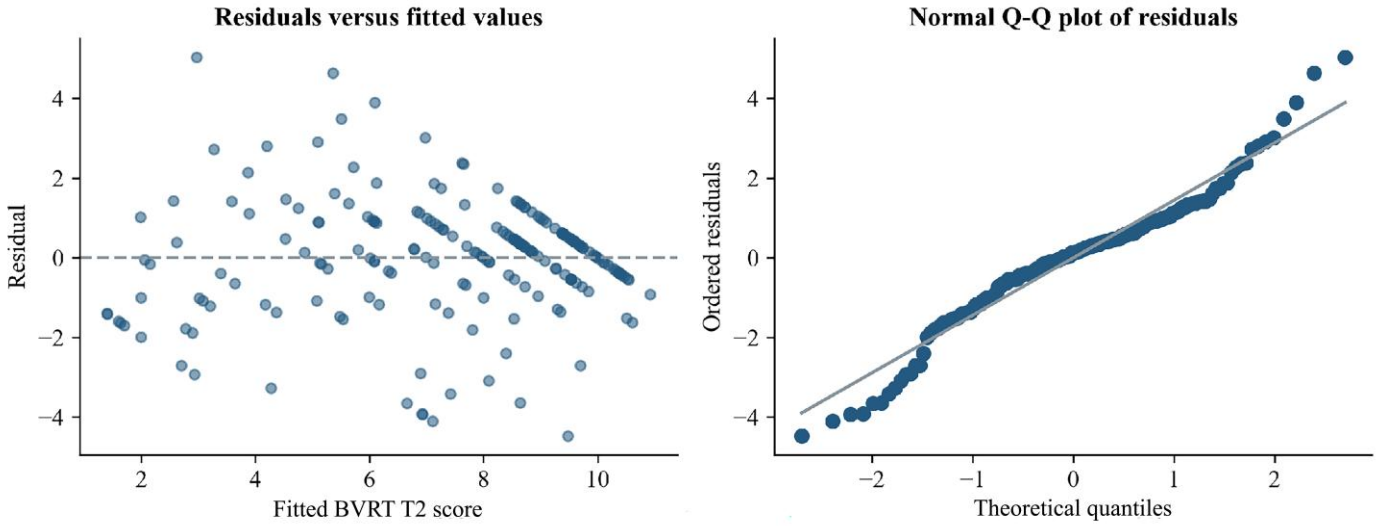


Figure 5. Residuals versus fitted BVRT T2 and the normal quantile plot for the adjusted linear model. Heteroscedasticity and fitted values above the scale prompted robust inference and the bounded-outcome sensitivity analysis.

In the fractional logit sensitivity model fitted to all 200 participants, PANSS N had a logit-scale coefficient of −0.026 (HC0 95% CI −0.068 to 0.016; $p = 0.229$). This likewise provided no clear evidence of an additional association with T2 after baseline adjustment (Table 7). The coefficients are on a different scale from the linear model and should not be compared as BVRT point differences.

Table 7. Fractional logit sensitivity model of BVRT T2 divided by ten.

Predictor or Contrast	Coefficient	HC0 95% CI	*p*
Intercept	−1.045	−2.227 to 0.137	0.083
PANSS N at T1 per point	−0.026	−0.068 to 0.016	0.229
BVRT T1 per point	0.466	0.407 to 0.524	<0.001
Age per year	−0.004	−0.021 to 0.012	0.593
Male versus female	0.125	−0.196 to 0.447	0.445
Higher versus secondary education	0.105	−0.233 to 0.443	0.544
Schizophrenia versus F23	0.264	−0.068 to 0.596	0.119

Note: n = 200. All observations, including scores of zero and ten, were retained. The same covariates as the main model were used. Coefficients and intervals are on the logit scale; *p* values use the normal distribution. The model is a sensitivity analysis and does not assume the ten card outcomes were available as independent trial-level records.

A post hoc extension of the linear model added (BVRT T1 minus its sample mean) squared. The quadratic coefficient was −0.034 ($p = 0.0106$); PANSS N remained imprecise (B = −0.029; 95% CI −0.084 to 0.025; $p = 0.290$). This check suggests that allowing curvature did not change the PANSS interpretation, while illustrating uncertainty about linear baseline adjustment. It was not used for model selection or clinical prediction.

Excluding the previously corrected total record left n = 199: mean change 0.839 (95% CI 0.621–1.057), adjusted PANSS N B = −0.031 (95% CI −0.087 to 0.026; $p = 0.287$). Excluding the three historical PANSS annotations left n = 197: mean change 0.843 (95% CI 0.623–1.062), PANSS N B = −0.038 (95% CI −0.095 to 0.019; $p = 0.193$). The origin of these records is described in the data-audit note (Section 2.5). The full 200-person analysis remains primary.

No records were excluded on the basis of the qualitative observations, and the legacy marker described in Section 2.5 was not used as an exclusion criterion.

3.7. Exploratory Diagnostic Comparisons

The mean change was 0.991 (95% CI 0.691–1.292) in schizophrenia (n = 113) and 0.632 (95% CI 0.320–0.944) in F23 (n = 87). The unadjusted difference was 0.359 (95% CI −0.071 to 0.789; Welch $p = 0.102$). The adjusted diagnosis coefficient from the main T2 model was 0.385 (95% CI −0.053 to 0.823; $p = 0.084$). Both tests yielded BH q = 0.102 within their two-test family. Their estimands differ: the first compares change means, and the second compares T2 conditional on equal observed baseline scores and covariates. Neither establishes equivalence. Table 8 and Figure 6 present the group-specific mean changes (generated using Python (Matplotlib 3.11.2). Group-specific mean changes were calculated from individual BVRT score differences (T2 minus T1), and 95% confidence intervals were estimated using Student's t-distribution; the results were displayed as point estimates with vertical confidence-interval bars.

Table 8. Mean change by diagnostic group.

Clinical Group	n	Mean Change	95% CI
Acute and transient psychotic disorders (F23)	87	0.632	0.320 to 0.944
Schizophrenia	113	0.991	0.691 to 1.292

Note: All 200 participants are represented. Change is T2 minus T1. Group-specific intervals describe each mean, not the between-group contrast. Schizophrenia minus F23: 0.359 points (95% CI −0.071 to 0.789; Welch $p = 0.102$). This contrast and the adjusted diagnosis coefficient formed a two-test BH family, with q = 0.102 for both. F23, acute and transient psychotic disorders.

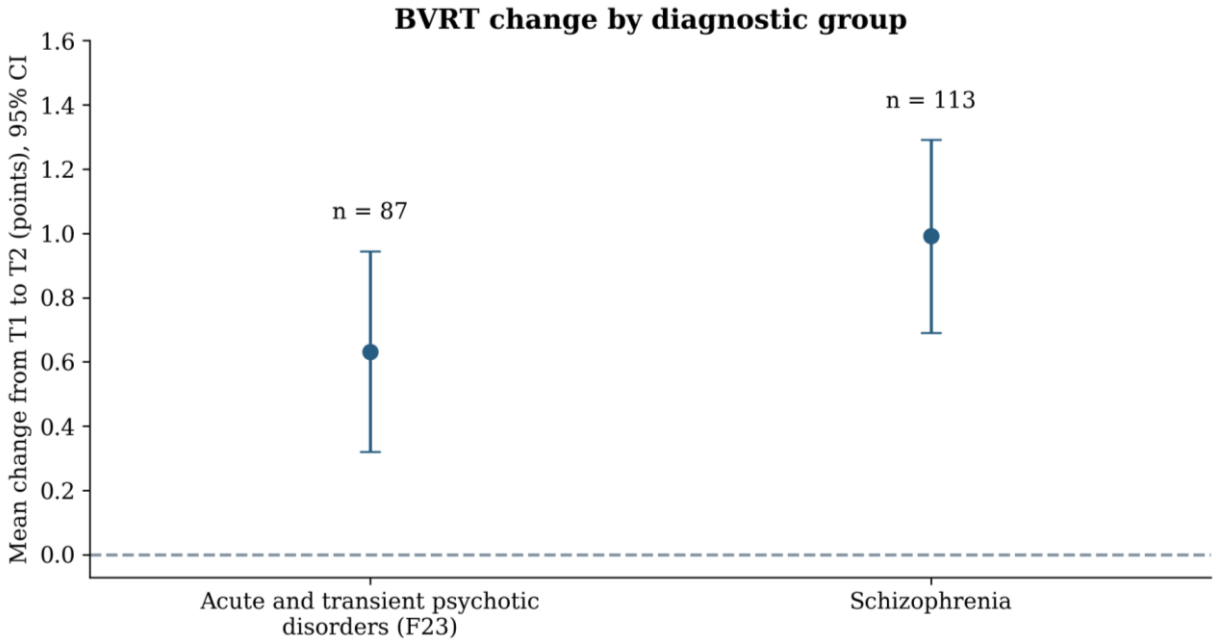


Figure 6. Mean within-participant change by diagnosis, with 95% confidence intervals for each group mean. These intervals do not represent the interval for the between-group difference, which is reported in the text and Table 8.

3.8. Exploratory Qualitative Drawing Observations

Recorded omissions, distortions, rotations and misplacements had lower mean counts at T2 after BH adjustment within the six-category paired t-test family (Table 9; Figure 7—generated using Python (Matplotlib 3.11.2); within-participant differences (T2

minus T1) were calculated separately for omissions, distortions, perseverations, rotations, misplacements, and size errors; mean changes and corresponding 95% confidence intervals were displayed as horizontal point estimates with confidence-interval bars; multiplicity-adjusted statistical tests were performed separately). Perseverations and size-error counts showed small mean changes with intervals spanning zero. The separate Wilcoxon family supported lower counts in the same four categories. These are separately coded qualitative observations, not a standardized BVRT error score. Multiple observations within one incorrect response do not create additional incorrect responses.

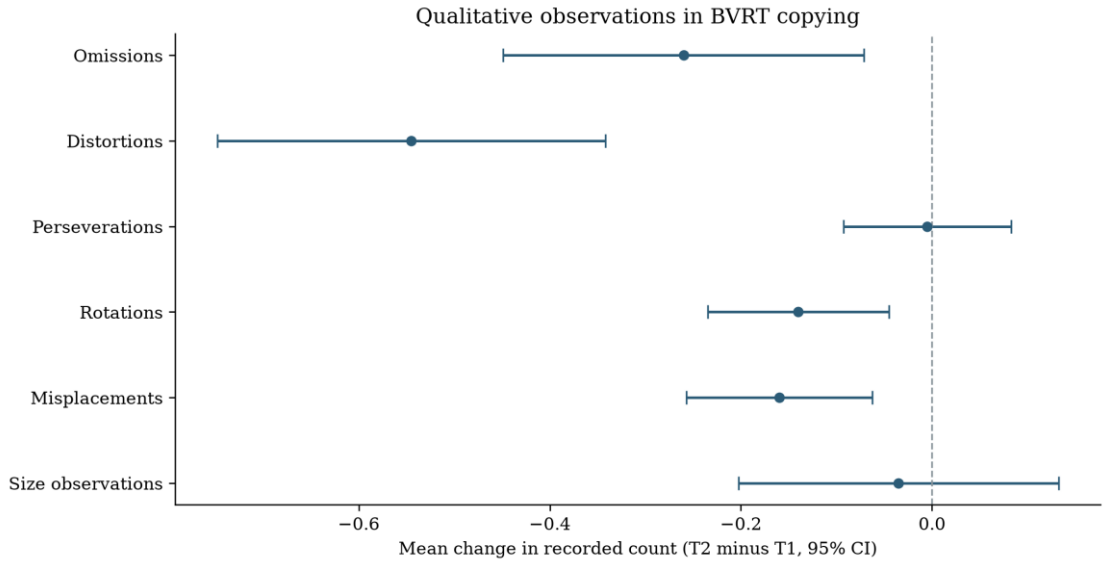


Figure 7. Mean change in separately coded qualitative observations, with individual 95% confidence intervals. Negative values indicate fewer recorded observations at T2. Binary correctness is unchanged by qualitative observations. Multiplicity-adjusted tests appear in Table 9.

The observations were unevenly distributed across participants (Table 9, Panel A). Distortions were the only category recorded in most participants at T1 (119/200, 59.5%), and the median count was 0 or 1 in every category. Perseverations were recorded in 19 participants (9.5%) at T1 and 10 (5.0%) at T2, and one participant accounted for 13 of the 25 perseverations recorded at T2. Misplacements were recorded in 38 participants (19.0%) at T1 and 15 (7.5%) at T2; counts decreased in 30 participants and increased in 6. Mean changes in these two categories therefore reflect a small subgroup of participants.

Table 9. Qualitative copying observations: distribution and paired changes.

Panel A. Distribution of counts across participants							
Category	**≥1 Observation T1, n (%)**	**≥1 Observation T2, n (%)**	**Median (IQR) T1**	**Median (IQR) T2**	**Range T1**	**Range T2**	**Decreased/Unchanged/Increased, n**
Omissions	65 (32.5)	55 (27.5)	0 (0–1)	0 (0–1)	0–11	0–12	42/139/19
Distortions	119 (59.5)	80 (40.0)	1 (0–2)	0 (0–1)	0–14	0–10	86/90/24
Perseverations	19 (9.5)	10 (5.0)	0 (0–0)	0 (0–0)	0–6	0–13	15/178/7
Rotations	57 (28.5)	35 (17.5)	0 (0–1)	0 (0–0)	0–4	0–3	40/143/17
Misplacements	38 (19.0)	15 (7.5)	0 (0–0)	0 (0–0)	0–5	0–5	30/164/6
Size errors	61 (30.5)	56 (28.0)	0 (0–1)	0 (0–1)	0–7	0–12	31/144/25
Panel B. Paired changes							

Category	T1 Mean (SD)	T2 Mean (SD)	Change 95% CI	t Test *p*	BH q	Wilcoxon BH q
Omissions	1.02 (2.12)	0.76 (1.77)	−0.260 −0.449 to −0.071	0.007	0.011	0.004
Distortions	1.64 (2.37)	1.10 (2.00)	−0.545 −0.748 to −0.342	<0.001	<0.001	<0.001
Perseverations	0.13 (0.53)	0.13 (0.97)	−0.005 −0.093 to 0.083	0.911	0.911	0.415
Rotations	0.38 (0.70)	0.24 (0.58)	−0.140 −0.235 to −0.045	0.004	0.008	0.004
Misplacements	0.28 (0.69)	0.12 (0.50)	−0.160 −0.257 to −0.063	0.001	0.004	<0.001
Size errors	0.66 (1.29)	0.62 (1.49)	−0.035 −0.202 to 0.132	0.681	0.817	0.415

Note: n = 200. Panel A shows the number (%) of participants with at least one recorded observation, the median (interquartile range) and range of counts at each visit, and the number of participants whose count decreased, remained unchanged or increased from T1 to T2. In Panel B, change = T2 minus T1; negative values indicate fewer recorded observations. Paired t and Wilcoxon tests each form a separate six-test BH family. CIs are individual t intervals for mean change. Wilcoxon assesses ranked differences, with zeros omitted, tie adjustment and no continuity correction. Statistical change in a category does not establish a specific cognitive mechanism or change in a formal error score.

Supplementary Table S1 reports the within-category comparisons (Panel A) and all 15 pairwise contrasts between category changes (Panel B). To examine how the qualitative observations related to the correct-card score, total qualitative counts were compared across correct-card scores (Supplementary Table S1, Panel C). The two measures were closely and inversely related (Spearman $\rho = -0.965$ at T1 and −0.927 at T2), as were their changes ($\rho = -0.786$). Every participant scoring below ten at T1 had at least one recorded observation, whereas none of the 25 participants with a maximum score at T1, and 2 of the 57 at T2, had any. Among the 58 participants whose correct-card score did not change, the total qualitative count changed in 14 (decreased in 9, increased in 5). In this sample, the qualitative counts therefore largely mirrored the binary correct-card score.

3.9. Exploratory Analysis of Longitudinal Differences Between Qualitative Observation Categories

A supplementary longitudinal Poisson GEE analysis suggested that the relative change in recorded qualitative drawing observations was not uniform across the six categories (time-by-category interaction: Wald $\chi^2(5) = 11.27$; $p = 0.0463$). This result was sensitive to the variance estimator: the interaction gave $\chi^2(5) = 11.16$ ($p = 0.048$) with a bias-reduced sandwich estimator and $\chi^2(5) = 10.40$ ($p = 0.065$) with a participant-level cluster-bootstrap covariance, so the evidence for between-category heterogeneity is weak. Category-specific T2/T1 time ratios (robust 95% CI) were 0.418 (0.233–0.750) for misplacements, 0.632 (0.460–0.867) for rotations, 0.668 (0.572–0.779) for distortions, 0.744 (0.603–0.917) for omissions, 0.947 (0.727–1.232) for size errors and 0.962 (0.475–1.947) for perseverations (Supplementary Table S2). Larger relative reductions were thus observed for misplacements, rotations, distortions and omissions, whereas size errors and perseverations changed relatively little; the wide interval for perseverations reflects the small number of participants with such observations.

In a separate exploratory analysis involving 15 pairwise comparisons of within-participant absolute changes, the reduction in distortion observations exceeded the

corresponding reductions in perseverations (difference between mean changes = −0.540; BH q = 0.00010), size errors (−0.510; q = 0.00040), rotations (−0.405; q = 0.00259), and misplacements (−0.385; q = 0.00585). These comparisons concern study-specific qualitative observation counts rather than standardized BVRT error scores. Differences in baseline category frequencies must also be considered when interpreting the magnitudes of these changes.

3.10. Exploratory Treatment and Symptom Associations

Each of the five indicators was added separately to the base T2 regression. A further model included auditory, visual and mixed hallucinations jointly, with no coded hallucinations as reference; its three terms were assessed by a robust global F test (3 and 190 df). Absence of a particular code in a separate-indicator model includes participants with other hallucination categories. These contrasts should not be read as comparisons with a uniformly asymptomatic group. Mean seven-day change was 0.819 cards (95% CI 0.544–1.093) in the incisive category (n = 138) and 0.871 cards (95% CI 0.519–1.223) in the sedative category (n = 62; Welch $p = 0.816$).

No treatment/symptom association met q < 0.05 within the six-test family (Table 10; Figure 8—generated using Python (Matplotlib 3.11.2) as a coefficient-and-confidence-interval plot; estimates were obtained from five separate ordinary least squares regression models, each incorporating one treatment- or symptom-related indicator alongside the base-model covariates; individual 95% confidence intervals were calculated using HC3 robust covariance estimates). Mixed hallucinations had a nominal inverse association with T2 (B = −0.484; 95% CI −0.936 to −0.031; $p = 0.036$), but its BH q was 0.218. The global hallucination test was also inconclusive (F(3, 190) = 1.877; $p = 0.135$; q = 0.405). These nonrandomised comparisons do not support comparative drug efficacy or a specific cognitive mechanism.

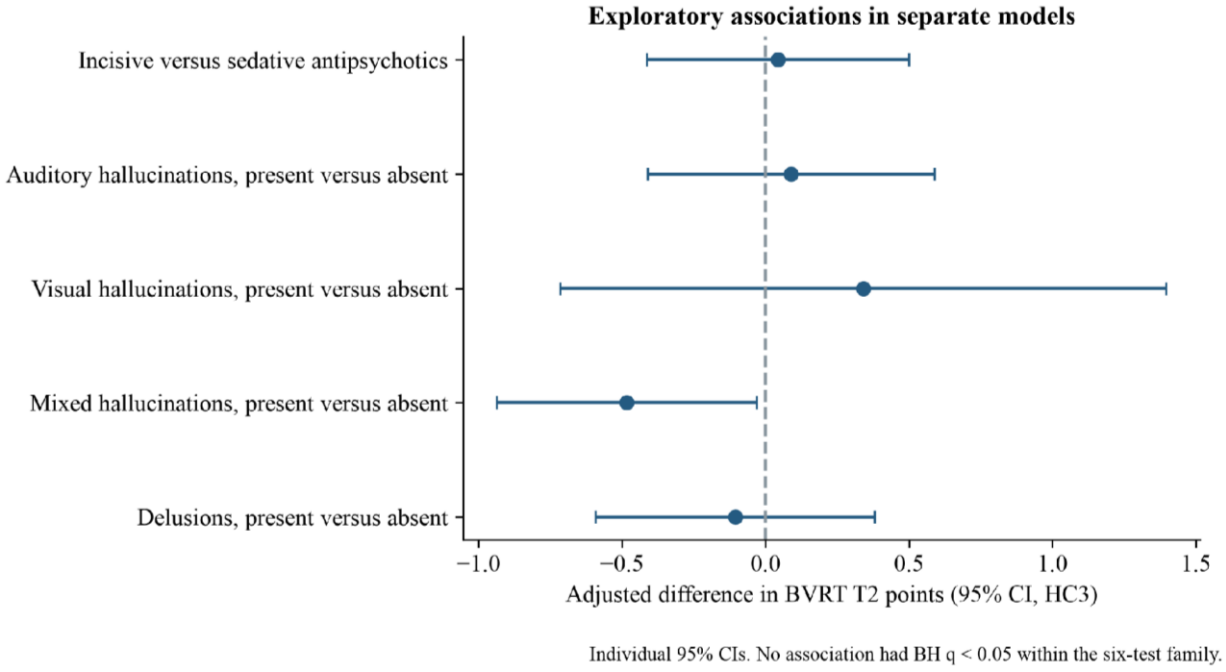


Figure 8. Associations from five separate adjusted T2 models, with individual HC3 95% confidence intervals. Each indicator was added to the base model. The source treatment labels are broad classifications; absence of a hallucination code can include other hallucination categories. Multiplicity-adjusted results and the global test are in Table 10.

Table 10. Exploratory associations with BVRT T2.

Exposure or Contrast	B	HC3 95% CI	*p*	BH q
Incisive versus sedative	0.044	−0.413 to 0.500	0.851	0.851
Auditory hallucinations	0.089	−0.411 to 0.589	0.725	0.851
Visual hallucinations	0.341	−0.714 to 1.397	0.524	0.851
Mixed hallucinations	−0.484	−0.936 to −0.031	0.036	0.218
Delusions	−0.104	−0.591 to 0.382	0.672	0.851
Hallucination categories global test	Not applicable	Not applicable	0.135	0.405

Note: n = 200 per model. The base model includes PANSS N, BVRT T1, age, sex, education and diagnosis. Global F = 1.877. BH adjustment covers the five indicator tests and the global test. Confidence intervals are individual intervals; a single coefficient and CI do not apply to the global test. Treatment classification is nonrandomised and lacks drug-specific adjustment.

3.11. Sensitivity of the Available Sample Size

For 200 pairs, minimum detectable standardized mean changes were dz = 0.199 at 80% power and 0.230 at 90% power. Assuming an SD of differences of 1.556, these correspond to 0.310 and 0.358 BVRT points. Under a separate equal-variance approximation for the diagnostic groups, 80% power required a between-group difference of approximately 0.622 points (Table 11). These are conditional sensitivity estimates, not thresholds of clinical importance.

Table 11. Minimum detectable differences under specified assumptions.

Contrast	n	Power	Standardized Difference	BVRT Points
Paired mean change	200	80%	0.199	0.310
Paired mean change	200	90%	0.230	0.358
Diagnostic-group change contrast	113/87	80%	0.402	0.622
Diagnostic-group change contrast	113/87	90%	0.465	0.720

Note: Paired differences use dz; group contrasts use the common-SD standardised difference. The observed mean effect was not used as a power target. Point conversions depend on the variance assumptions and are not validated minimum clinically important differences. These estimates do not imply that nonsignificant results demonstrate no association or that the sample is adequate for every possible hypothesis.

For the completed study, confidence intervals describe precision directly. The primary mean-change CI was 0.618–1.052 BVRT points, whereas the diagnostic difference CI was −0.071 to 0.789. The adjusted PANSS N interval was −0.087 to 0.026 points per PANSS point. Sample-size sensitivity should be read alongside these intervals and the substantial design and measurement constraints, rather than used to reinterpret a nonsignificant test as evidence of equivalence.

4. Discussion

4.1. Meaning of the Observed Change

Copying performance increased by an average of 0.835 correctly copied cards over seven days in this consecutive inpatient sample. Baseline PANSS N accompanied poorer concurrent performance, but its additional association with retest performance was uncertain after baseline adjustment. The primary result is an estimate of observed change under a repeated copying schedule. It does not identify which clinical or measurement processes produced that change.

The same cards were shown twice, making familiarity with the designs and instructions a plausible contributor. Retest effects vary with task, interval, population and use of alternate forms [19]. They are typically largest when identical forms are readministered after short intervals [19], as in the present design. Because the stimulus remains visible during copying, improvement need not depend on remembering the designs; familiarity with the figures, the instructions and an efficient drawing strategy may suffice [92]. Schizophrenia research has repeatedly demonstrated this problem: Goldberg et al. found cognitive gains after antipsychotic treatment similar to those in repeatedly tested healthy participants [93,94], and Keefe et al. found prebaseline gains comparable to gains during placebo exposure in cognition trials [95,96]. These findings support caution but do not quantify the practice component of this BVRT result. Without an alternate-form or non-retested comparison group, the relative contributions of practice and clinical stabilization to the mean increase of 0.835 cards cannot be estimated; counterbalanced alternate BVRT forms or a retested control group would be required to separate them.

Clinical stabilization, treatment, task engagement and measurement variation may also contribute. The paired effect size describes change relative to the variability of individual differences, rather than efficacy against a control condition. Likewise, participants with higher scores cannot be classified as clinical responders without reliable-change estimates and an external criterion of importance [97]. Copying under continuous visual presentation does not establish improved visual memory, and a single copying score does not provide the broader profile addressed by domain-based cognitive assessment [2,98].

Antipsychotic initiation and dose adjustment during the first days of hospitalization coincided with the observation interval and may influence copying in either direction, through symptom reduction and improved engagement or through sedation and extrapyramidal or anticholinergic effects on fine motor performance [21]. Doses, dose changes and as-needed interventions between T1 and T2 were not recorded; the only available indicator, the broad treatment category at T1, was not associated with change (Section 3.10), but it cannot capture these influences.

4.2. Clinical Associations and Diagnostic Uncertainty

The inverse association between conventional PANSS negative-subscale severity and concurrent BVRT copying performance remained evident after adjustment for demographic characteristics, diagnosis, and the positive and general psychopathology PANSS subscales. Consistent with previous evidence linking negative symptom severity to cognitive impairment in schizophrenia [24,99], this finding supports an association between the clinical dimension represented by PANSS N and visuoconstructive task performance that is not entirely explained by the other measured symptom dimensions. However, the conventional PANSS negative subscale includes items reflecting cognitive and disorganization-related characteristics, particularly N5 (difficulty in abstract thinking) and N7 (stereotyped thinking), which are not considered core negative symptoms in contemporary assessment frameworks [18]. In factor-analytic models of the PANSS, N5 loads on a disorganized/concrete factor together with conceptual disorganization and poor attention, and N7 is not retained on the negative factor. These items share cognitive demands with visuoconstructive copying, such as abstraction, flexible organization of the figure and planning, so part of the observed association may reflect cognitive and disorganization features rather than diminished expression or motivation. The post hoc anergia analysis suggests that the association is not attributable solely to these items, because the anergia cluster, which excludes them, was also associated with poorer copying; however, PANSS N retained an association after adjustment for anergia, so the cognitive items may contribute to it (Section 3.5). Item-level data, the consensus five-factor PANSS model or dedicated negative-symptom instruments would be needed to separate these contributions.

Consequently, the observed relationship cannot be attributed specifically to primary negative symptoms. Furthermore, the absence of a clearly demonstrated additional association between baseline PANSS N and retest performance after adjustment for initial BVRT scores indicates that the present findings should not be interpreted as evidence that negative symptoms independently predict short-term performance change. Item-level PANSS data and repeated symptom assessments would be necessary to investigate the underlying clinical relationships more precisely.

The adjusted T2 model asks a different question from the baseline correlation: it compares retest scores among participants with equal observed initial performance and covariates. Its CI includes both a potentially relevant negative association and a small positive association. Attenuation after baseline adjustment does not establish that negative symptoms are irrelevant or that baseline performance mediates their influence, as changes in adjusted associations depend on the underlying causal structure and do not, by themselves, demonstrate mediation [100,101]. Measurement error and unmeasured clinical differences remain possible explanations.

The diagnostic contrast was also imprecise. F23 participants had lower symptom scores and higher baseline copying scores, leaving different room for improvement. A longitudinal study of first-episode F23 and schizophrenia reported no statistically significant difference in cognitive trajectories, but used different domains, selected F23 subtypes and a much longer interval [102]. Its findings therefore provide context rather than direct replication. F23 diagnoses may also follow heterogeneous subsequent courses [41]. The present nonsignificant contrast does not establish equivalent trajectories, particularly given the more limited sensitivity for between-group differences.

4.3. Differential Patterns of Qualitative Copying Observations

Beyond the increase in the number of correctly copied cards, recorded qualitative drawing observations decreased in four of the six categories (distortions, omissions, rotations and misplacements), whereas perseverations and size errors changed little. The joint longitudinal analysis provided only weak evidence that relative changes differed between categories: the interaction was nominally significant with the robust sandwich estimator but not with a cluster-bootstrap covariance ($p = 0.065$). Pairwise comparisons indicated larger absolute reductions in distortions than in several other observation types, but distortions were also the most frequent category at baseline. Moreover, total qualitative counts closely mirrored the correct-card score ($\varrho = -0.965$ at T1), observations were almost absent among participants with maximum scores, and changes in the two measures were strongly related. These results show that the change in correct-card scores was accompanied by changes in the recorded characteristics of the drawings, but they do not show that qualitative observations add information beyond the correct-card score; whether they do remains a question for future validation.

Previous investigations have documented differences between specific BVRT error categories in other populations, including longitudinal age-related changes, and qualitative abnormalities in BVRT performance have also been reported in schizophrenia [11,13]. The potential contribution of the present study therefore lies not in identifying previously unknown error types, but in providing a short-interval longitudinal characterization of qualitative copying observations in a hospitalized sample with schizophrenia or acute and transient psychotic disorders.

These findings must nevertheless be interpreted cautiously. Differences in initial observation frequencies complicate comparisons between absolute changes, and changes in the sparse categories (perseverations and misplacements) were driven by small subgroups of participants. The study-specific qualitative observations were coded retrospectively from the completed drawings, were not equivalent to a standardized BVRT Number Error

Score, and were not supported by independently verified interrater reliability estimates. Accordingly, the findings do not establish selective recovery of visuospatial processing, executive control, or any specific cognitive mechanism. Whether qualitative assessment provides information beyond global copying accuracy should be examined in future studies using standardized scoring, verified interrater reliability, external criteria and appropriate retest comparison conditions. Broad treatment categories and unmeasured medication doses also prevent conclusions about comparative pharmacological effects.

4.4. Comparison with Recent Studies (2023–2025)

Recent research has provided further evidence regarding visuospatial and cognitive impairment in psychotic disorders, the relationship between psychopathological symptom severity and cognitive performance, and changes observed during repeated neuropsychological assessment. Nevertheless, studies specifically examining short-interval changes in BVRT copying accuracy and qualitative drawing observations in patients with schizophrenia and acute and transient psychotic disorders (ATPD) remain limited. The present findings should therefore be interpreted in relation to studies employing comparable cognitive tasks or clinical populations, while acknowledging differences in assessment procedures, diagnostic composition, and follow-up intervals.

One of the most directly relevant investigations was conducted by Stoychev et al. (2025) [11], who compared cognitive performance and symptom severity in 62 patients with schizophrenia or substance-induced psychotic disorders, using a neuropsychological battery that included the Benton Visual Retention Test. Patients with schizophrenia demonstrated poorer BVRT performance, including fewer correct responses and greater frequencies of specific error types, particularly distortions, perseverations, and misplacements. These findings support the relevance of examining qualitative response characteristics alongside quantitative BVRT performance in psychotic disorders. In our sample, changes in copying accuracy were accompanied by reductions in recorded qualitative observations, particularly distortions, omissions, rotations and misplacements, although these counts closely mirrored the correct-card score and evidence for differences between categories was weak. However, Stoychev et al. employed a cross-sectional design and investigated a different diagnostic comparison. Moreover, differences in BVRT administration and scoring procedures limit direct comparisons of the quantitative and qualitative outcomes. Our study extends this area of research by examining the same participants at two assessments separated by seven days, although the exploratory nature and study-specific recording of qualitative observations require cautious interpretation.

Regarding diagnostic comparisons, Knížková et al. (2024) [40] investigated the longitudinal course of cognitive functioning in 70 patients with first-episode schizophrenia or ATPD, including 39 participants with schizophrenia and 31 with ATPD. Cognitive performance was evaluated over approximately 12 months, focusing on processing speed, working memory, and executive functioning. Although both diagnostic groups demonstrated improvements across the examined cognitive domains, no statistically significant differences in their longitudinal cognitive trajectories were identified. These observations are broadly consistent with the inconclusive diagnostic-group comparisons in our study, in which the seven-day change in BVRT copying accuracy did not differ significantly between schizophrenia and ATPD after correction for multiple comparisons. However, the two investigations differ substantially in assessment interval, cognitive domains, and patient selection. In particular, Knížková et al. examined selected first-episode ATPD presentations, whereas the present F23 sample was not restricted to first-episode psychosis. Consequently, the absence of statistically significant diagnostic contrasts in either study should not be interpreted as demonstrating equivalent cognitive trajectories.

Additional research concerning cognitive changes during psychiatric hospitalization was provided by Pirani et al. (2025) [14], who evaluated 84 hospitalized patients with schizophrenia or bipolar disorder with psychotic features using the Clock Drawing Test and psychopathological assessments at admission and discharge. Both groups demonstrated improved drawing-test performance, with greater changes observed in the bipolar disorder group. The authors did not identify significant correlations between changes in Clock Drawing Test performance and changes in positive or negative symptoms. Their findings are consistent with the possibility that drawing-based cognitive performance can improve during hospitalization without corresponding closely to changes in psychopathological symptom severity. In our study, BVRT copying accuracy increased over seven days, whereas baseline PANSS negative-subscale scores did not demonstrate a clear additional association with follow-up performance after adjustment for baseline BVRT scores and relevant covariates. Nevertheless, the Clock Drawing Test and BVRT copying administration assess overlapping but nonidentical cognitive processes, and the clinical populations and assessment intervals also differ. Therefore, improvement observed using these instruments cannot be assumed to reflect the same underlying cognitive changes.

The importance of accounting for repeated-testing effects is further illustrated by Arnautovska et al. (2023) [103], who examined verbal and visual learning in 86 patients with treatment-resistant schizophrenia receiving clozapine. Participants underwent repeated neuropsychological assessments at baseline and at weeks 8, 24, and 52. No statistically significant improvements in verbal or visual learning were detected across the assessment period. Although these findings differ from the increased BVRT copying accuracy observed in our sample, the discrepancy may reflect differences in clinical characteristics, cognitive domains, test administration, and assessment intervals. In particular, the Brief Visuospatial Memory Test–Revised used by Arnautovska et al. assesses visual learning and memory, whereas BVRT Administration C in the present investigation involves copying continuously visible stimuli. These studies collectively illustrate that performance changes during repeated cognitive assessment are task- and population-dependent. Moreover, the absence of an alternate-form or non-retested comparison condition in our study prevents separation of practice-related improvement from potential clinical influences.

The relationship between negative symptoms and cognitive functioning has also received considerable attention in recent research. Kalisova et al. (2023) [104] investigated 211 patients with schizophrenia and demonstrated that cognitive impairment and negative symptom severity were associated with psychosocial functioning, highlighting the importance of considering these clinical dimensions together. Similarly, Blotner et al. (2025) [105], in a longitudinal investigation of 67 individuals with early psychosis, reported associations between negative symptom measures and cognitive performance, while identifying negative symptom dimensions related to motivation and pleasure as predictors of subsequent functioning. These findings provide clinical context for the inverse association between baseline PANSS negative-subscale severity and concurrent BVRT copying accuracy observed in our study. This association remained evident after adjustment for demographic variables, diagnosis, and the PANSS positive and general psychopathology dimensions. However, the conventional PANSS negative subscale includes cognitive and disorganization-related items, and the relationship observed in our study cannot be attributed specifically to core negative symptoms. The lack of a clearly demonstrated additional association with adjusted retest performance indicates that concurrent symptom–cognition relationships should be distinguished from longitudinal prediction of performance change. The use of different negative-symptom instruments and cognitive outcomes across studies further limits direct comparison.

Recent investigations support the clinical relevance of assessing visuospatial and dra ing-based performance, negative symptom severity, and longitudinal cognitive changes in psychotic disorders. The present study contributes complementary descriptive evidence by combining seven-day repeated BVRT copying scores with separately coded qualitative drawing observations in a relatively large hospitalized sample comprising patients with schizophrenia and ATPD. Its principal contribution lies in characterizing short-interval changes in copying accuracy and exploring differences among qualitative observation trajectories, rather than demonstrating treatment-related cognitive recovery or disorder-specific cognitive improvement. The findings also underscore the need for future longitudinal investigations incorporating standardized qualitative scoring procedures, independent reliability assessment, alternate test forms or appropriate retest controls, repeated psychopathological assessments, and longer follow-up intervals to determine the reproducibility and clinical significance of the observed changes.

4.5. Limitations and Implications

The absence of a control condition and the repeated administration of identical BVRT stimuli introduce potential testing-related bias and temporal confounding. As described in the foundational work of Campbell and Stanley (1963) cited in 2012 by Seel [106], repeated-measures designs without an appropriate comparison condition are vulnerable to testing, history, and maturation effects. In the present study, familiarity with the stimulus cards, task instructions, and copying procedures may have contributed to improved performance independently of genuine changes in visuoconstructive ability. Contemporary neuropsychological research continues to emphasize the importance of accounting for practice effects when interpreting repeated cognitive assessments [107]. Moreover, treatment initiation, clinical stabilization, and fluctuations in attention, motivation, or psychomotor functioning occurred within the same observation period. Doses, dose changes and acute or as-needed pharmacological interventions between the two assessments were not recorded, so pharmacological influences on the observed change could not be examined beyond the broad treatment category. These concurrent influences introduce potential time-related confounding, making it impossible to attribute the observed mean improvement of 0.835 correctly copied cards specifically to treatment or cognitive recovery. As emphasized by Dahabreh and Bibbins-Domingo (2024), causal interpretations of observational findings require careful consideration of uncontrolled confounding and the assumptions necessary for causal identification [108].

The short assessment interval and the availability of only two measurements limit the interpretation of longitudinal change and introduce uncertainty arising from measurement variability. Although the seven-day interval enabled characterization of short-term copying performance, it was insufficient to establish the persistence of improvement, distinguish temporary fluctuations from sustained change, or determine clinically meaningful individual recovery. Jacobson and Truax (1991) [109] emphasized the importance of distinguishing statistically detectable differences from reliable and clinically significant change. Karr et al. (2024) demonstrated the value of test–retest reliability and practice-adjusted reliable-change estimates for interpreting individual neuropsychological changes [107]. Because comparable reliable-change reference values and external functional outcomes were unavailable for the present BVRT protocol, the observed group-level improvement cannot be interpreted as evidence of reliable individual cognitive recovery. As Nuechterlein et al. (2025) highlighted, cognitive test performance and everyday functional outcomes represent related but distinct dimensions of clinical assessment in schizophrenia [110].

The bounded nature of the BVRT correct-card score introduces potential ceiling effects, while analyses relating baseline scores to subsequent changes are susceptible to

regression to the mean and mathematical coupling. The correct-card score ranges from zero to ten, restricting the potential for improvement among participants with high initial scores. In the present sample, the proportion of participants achieving the maximum score increased between T1 and T2. Specifically, 25 participants (12.5%) scored ten at T1 and could not show improvement, a further 35 reached ten at T2 so that any additional gain could not be registered, and 95 (47.5%) scored nine or ten at T2. The full-sample mean change of 0.835 cards is therefore likely to underestimate the change that a less constrained measure would have shown among participants with lower initial scores (0.971 cards after excluding participants starting at ten; 1.122 cards among those scoring below nine), although these subgroup estimates are also affected by regression to the mean. The ceiling also reduces the variance of retest scores and thus the sensitivity of analyses of retest performance, including the adjusted PANSS N and diagnostic comparisons; more F23 than schizophrenia participants scored ten at baseline (17.2% vs. 8.8%). In contrast, the concurrent PANSS N association was not an artefact of the ceiling (Spearman $\varrho = -0.281$ after excluding participants scoring ten at T1). Tasks with greater difficulty or finer-grained scoring would be needed to measure short-term change across the full range of performance. Consequently, ceiling effects may have reduced the sensitivity of the outcome to further improvement, particularly among initially high-performing participants. Oldham (1962) [111] demonstrated that associations between baseline measurements and change scores may be distorted by mathematical dependence, while Barnett et al. (2005) [112] described how regression to the mean can complicate the interpretation of repeated measurements. Recent neuropsychological research continues to demonstrate the influence of ceiling effects on cognitive assessment and longitudinal score interpretation [113]. Therefore, the larger observed gains among participants with lower baseline BVRT scores should not be interpreted as evidence of greater treatment responsiveness or preferential cognitive recovery.

The use of ordinary least squares regression for a bounded outcome introduces potential model misspecification and heteroscedasticity-related uncertainty. Although heteroscedasticity-consistent HC3 covariance estimates were used, robust standard errors do not resolve incorrect functional-form assumptions or prevent predicted values from exceeding the permissible score range. The fractional-response framework developed by Papke and Wooldridge [114] provides an alternative for outcomes bounded between zero and one. Its continued relevance to bounded health outcomes has been emphasized by Mullahy (2024) [69]. Accordingly, a fractional-logit sensitivity analysis was performed using the BVRT T2 score divided by ten. However, agreement between alternative analytical approaches cannot eliminate potential measurement error or fully establish the adequacy of the statistical models. The resulting associations should therefore be interpreted cautiously.

Limitations in the documentation and independent verification of BVRT scoring introduce potential measurement and observer-related bias. Although the same copying administration procedure and scoring approach were used at both assessments, detailed scoring tolerance rules and independently verified interrater reliability estimates were unavailable. The study-specific qualitative drawing observations were recorded separately from binary correct-card classification and were not equivalent to a standardized BVRT Number Error Score. Because they were coded retrospectively from the completed drawings, intermediate drawing behaviour and spontaneous corrections were not captured, and the close correspondence between total observation counts and incorrect cards indicates that the two measures largely overlapped. The observations were coded by a single psychologist who was unaware of PANSS scores but aware of the diagnoses and not blinded to assessment visit, because the T1 and T2 drawings were available together; no second rater coded the drawings, so agreement could not be monitored, and expectations

of improvement could have influenced the classification of the T2 drawings. These circumstances introduce uncertainty regarding measurement consistency and comparability with studies using conventional BVRT error-scoring procedures. Cohen established a foundational approach to quantifying agreement between observers [115], while recent neuropsychological research continues to emphasize the importance of standardized assessment and interrater consistency [116]. Without independently established scoring reliability, the possibility that some observed changes in qualitative drawing characteristics reflect classification variability rather than changes in underlying performance cannot be excluded.

The exploratory and post hoc analyses introduce potential multiplicity-related bias, selective-inference concerns, and an increased risk of false-positive findings. The supplementary longitudinal and multivariable analyses were not prospectively prespecified, and several families of statistical comparisons were investigated. Simmons et al. (2011) demonstrated that analytical flexibility can substantially increase the probability of obtaining apparently significant findings [117]. Recent methodological research has similarly emphasized the susceptibility of post hoc subgroup and interaction analyses to inflated type I error and inappropriate interpretation [118]. Although the Benjamini–Hochberg procedure was applied within defined families of exploratory comparisons, this approach did not eliminate uncertainty associated with the broader collection of analyses or the absence of documented prospective analytical prespecification. In particular, the time-by-category interaction reached nominal significance only with the robust sandwich estimator (cluster-bootstrap $p = 0.065$); it should be regarded as hypothesis-generating and requires independent replication.

Recruitment from a single psychiatric hospital and the study-specific eligibility requirements introduce potential selection bias and limit external validity. Although consecutive recruitment reduced investigator discretion in identifying eligible participants, the analyzed sample was restricted to hospitalized adults meeting specific educational criteria. Hospital-based sampling can produce selection-related distortions when inclusion depends on clinical or demographic characteristics associated with the outcomes of interest, an issue originally discussed by Berkson [119]. Recent schizophrenia research continues to highlight the challenges of obtaining representative psychiatric research samples and the influence of recruitment procedures on study participation [120]. Therefore, the findings may not be directly generalizable to community-dwelling patients, individuals with lower educational attainment, or populations treated in different clinical settings.

Diagnostic heterogeneity and limitations in psychopathological assessment introduce potential classification and construct-measurement bias. Acute and transient psychotic disorders encompass heterogeneous clinical presentations and may follow different subsequent diagnostic trajectories. The ICD-10 F23 category comprises subtypes with and without symptoms of schizophrenia and predominantly delusional presentations; subtypes were not recorded, and the F23 group was not restricted to first episodes. Because a substantial proportion of individuals with acute and transient or brief psychotic disorders are later rediagnosed with schizophrenia-spectrum or affective disorders [30], the F23 group probably combined patients with remitting courses and patients in an early phase of a more persistent disorder. Diagnoses at follow-up were not available, so outcome-based subgroups could not be compared. This heterogeneity would tend to blur differences between the diagnostic groups and limits the generalization of the F23 results to any particular clinical course; the diagnostic comparisons describe groups defined by the admission diagnosis rather than distinct disorders. The consequences of classification errors for statistical associations were described in foundational methodological research by Bross [121], while Knížková et al. (2024) demonstrated the clinical and cognitive heterogeneity relevant to longitudinal comparisons between ATPD and schizophrenia [40]. In

addition, PANSS was administered only at baseline, preventing evaluation of whether changes in symptom severity accompanied changes in BVRT performance. The conventional PANSS negative subscale also includes cognitive and disorganization-related components and does not comprehensively capture all contemporary negative-symptom domains. A recent systematic evaluation of PANSS measurement properties identified important limitations in its content and structural validity [48]. Consequently, the observed association between PANSS N and baseline copying accuracy cannot be attributed specifically to primary negative symptoms, nor can the present data establish whether changes in psychopathology influenced subsequent copying performance.

Despite these limitations, the common assessment interval, complete paired measurements among analyzed participants, and explicit distinction between copying accuracy and qualitative observations support the descriptive contribution of the study.

5. Conclusions

The present observational study aimed to investigate seven-day changes in Benton Visual Retention Test (BVRT) copying performance and their associations with baseline negative symptom severity in hospitalized patients with schizophrenia or acute and transient psychotic disorders. The results supported both principal study hypotheses: first, that BVRT copying accuracy would increase over the seven-day assessment interval, and second, that greater baseline PANSS negative-subscale severity would be associated with poorer concurrent copying performance. Specifically, among 200 participants, the number of correctly copied cards increased by a mean of 0.835 points (95% CI: 0.618–1.052; $p < 0.001$), while baseline PANSS negative-subscale scores were inversely correlated with copying accuracy (Spearman's $\rho = -0.313$; $p < 0.001$). The latter association remained statistically significant after adjustment for positive and general psychopathology dimensions and demographic and diagnostic covariates. However, baseline negative symptom severity did not demonstrate a clear additional association with seven-day retest performance after adjustment for initial BVRT scores and relevant covariates (B = −0.031; 95% CI: −0.087 to 0.026; $p = 0.286$). Recorded qualitative drawing observations decreased in parallel with the correct-card gains, but evidence for differences between observation categories was weak, and diagnostic-group differences remained inconclusive.

These findings extend the descriptive evidence on short-interval visuoconstructive task performance during hospitalization for psychotic disorders. Because the qualitative counts largely mirrored the correct-card score, whether qualitative drawing characteristics add information to global accuracy scores remains a question for future validation. However, the observed changes cannot be disentangled from task familiarity, clinical stabilization, and other concurrent influences. The results do not establish treatment-related cognitive improvement, enhanced visual memory, differential diagnostic recovery, or clinically meaningful individual change. Standardized qualitative scoring, appropriate retest comparison conditions, repeated clinical assessments, and independent replication are needed to determine whether the observed patterns have reproducible neuropsychological or clinical significance.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/jcm15239201/s1, Table S1: Exploratory longitudinal comparisons of qualitative BVRT observation categories; Table S2: Poisson generalized estimating equations model of qualitative observation counts; File S1: Data dictionary; File S2: Python analysis code.

Author Contributions: Conceptualization, V.M. and S.T.; methodology, V.M.; software, A.G.Z.; validation, V.M., A.G.Z. and S.T.; formal analysis, V.M.; resources, S.T.; data curation, A.G.Z. and S.T.;

writing—original draft preparation, V.M. and A.G.Z.; writing—review and editing, V.M. and S.T.; supervision, S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the Doctoral School at Carol Davila University of Medicine and Pharmacy, Bucharest (Approval Code: 34990; Approval Date: 5 November 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patient(s) to publish this paper.

Data Availability Statement: The de-identified participant-level dataset is not publicly available because participants did not consent to public data sharing, but it is available from the corresponding author on reasonable request. The data dictionary and the Python analysis code are provided as Supplementary Materials (Files S1 and S2).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lee, M.; Cernvall, M.; Borg, J.; Plavén-Sigray, P.; Larsson, C.; Erhardt, S.; Sellgren, C.M.; Fatouros-Bergman, H.; Cervenka, S. Cognitive Function and Variability in Antipsychotic Drug–Naive Patients with First-Episode Psychosis: A Systematic Review and Meta-Analysis. *JAMA Psychiatry* **2024**, *81*, 468. https://doi.org/10.1001/jamapsychiatry.2024.0016.
2. Nuechterlein, K.H.; Green, M.F.; Kern, R.S.; Baade, L.E.; Barch, D.M.; Cohen, J.D.; Essock, S.; Fenton, W.S.; Frese, F.J.; Gold, J.M.; et al. The MATRICS Consensus Cognitive Battery, Part 1: Test Selection, Reliability, and Validity. *Am. J. Psychiatry* **2008**, *165*, 203–213. https://doi.org/10.1176/appi.ajp.2007.07010042.
3. Nuechterlein, K.H.; Green, M.F.; Kern, R.S. The MATRICS Consensus Cognitive Battery: An Update. In *Cognitive Functioning in Schizophrenia: Leveraging the RDoC Framework*; Barch, D.M., Young, J.W., Eds.; Current Topics in Behavioral Neurosciences; Springer International Publishing: Cham, Switzerland, 2022; Volume 63, pp. 1–18.
4. Diaz-Orueta, U.; Rogers, B.M.; Blanco-Campal, A.; Burke, T. The Challenge of Neuropsychological Assessment of Visual/Visuo-Spatial Memory: A Critical, Historical Review, and Lessons for the Present and Future. *Front. Psychol.* **2022**, *13*, 962025. https://doi.org/10.3389/fpsyg.2022.962025.
5. Silvan, A.B. *Benton Visual Retention Test*, 5th ed.; The Psychological Corporation: San Antonio, TX, USA, 1992.
6. Östling, A.-K.; Dahl, B.; Andersson, G.; Björck, C. Self-Assessed Memory Problems in Everyday Life in Relation to Assessed Cognitive Domains in Neuropsychological Tests: A Longitudinal Study of Adult Patients with Acquired Brain Injury. *Appl. Neuropsychol., Adult* **2026**, 1–13. https://doi.org/10.1080/23279095.2026.2734319.
7. Seo, E.H.; Lee, D.Y.; Choo, I.H.; Youn, J.C.; Kim, K.W.; Jhoo, J.H.; Suh, K.W.; Paek, Y.S.; Jun, Y.H.; Woo, J.I. Performance on the Benton Visual Retention Test in an Educationally Diverse Elderly Population. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* **2007**, *62*, P191–P193. https://doi.org/10.1093/geronb/62.3.P191.
8. Ceresetti, R.; Celle, S.; Roche, F.; Barthélémy, J.-C.; Michael, G.A.; Borg, C. Normative Data for a Neuropsychological Test Battery in the French Aging Population: 20-Year Follow-Up from the "PROgnostic OF Cardiovascular and Cerebrovascular Events" Study. *Arch. Clin. Neuropsychol.* **2025**, *40*, 467–485. https://doi.org/10.1093/arclin/acad074.
9. Moses, J.A. Factor Structure of Benton's Tests of Visual Retention, Visual Construction, and Visual Form Discrimination. *Arch. Clin. Neuropsychol.* **1986**, *1*, 147–156. https://doi.org/10.1093/arclin/1.2.147.
10. Simsarian, C.; Baldini, D.; Moses, J.A.; Rader, C.; McGrath, H.M.; Hammond, B.E.; Bongar, B. Exploratory Principal Component Analysis of the Benton Visual Form Discrimination Test, Matching Subtest (VFDT-MAT) Using the WAIS-IV, Visuospatial, and Language Tasks. *Appl. Neuropsychol. Adult* **2026**, 1–14. https://doi.org/10.1080/23279095.2026.2635554.
11. Stoychev, K.; Veleva, I.; Mineva-Dimitrova, E.; Chumpalova, P.; Dimitrova, E. Symptom Severity and Cognitive Performance in Patients with Substance Induced Psychotic Disorder and Schizophrenia: A Cross-Sectional Comparative Study. *Schizophr. Res. Cogn.* **2025**, *42*, 100388. https://doi.org/10.1016/j.scog.2025.100388.
12. Di Cecca, A.; Ilardi, C.R.; Della Pia, F.; Criscuolo, C.; Della Sala, S.; Salvatore, E. Distortion Errors Characterise Visuo-Constructive Performance in Huntington's Disease. *Clin. Neuropsychol.* **2025**, *39*, 1615–1633. https://doi.org/10.1080/13854046.2024.2411740.

13. Resnick, S.M.; Trotman, K.M.; Kawas, C.; Zonderman, A.B. Age-Associated Changes in Specific Errors on the Benton Visual Retention Test. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* **1995**, *50B*, P171–P178. https://doi.org/10.1093/geronb/50B.3.P171.
14. Pirani, A.; Sarabi-Jamab, A.; Bostan, F.; Salehian, R.; Ilaghi, M.; Aledavoud, A.; Shariat, S.V.; Mirfazeli, F.S. Differential Cognitive and Clinical Improvements in Schizophrenia and Bipolar Disorder Following Hospitalization: A Comparative Analysis Based on the Clock Drawing Test. *PLoS ONE* **2025**, *20*, e0325537. https://doi.org/10.1371/journal.pone.0325537.
15. Szczepańska, M.; Twardochleb, Z.; Miś, M.; Miś, M.; Druszcz, A.; Paprocka-Borowicz, M.; Rosińczuk, J. Effect of Transcranial Direct Current Stimulation on Memory and Emotional Recovery in Patients with Stroke and Traumatic Brain Injury: A Prospective, Multicenter, Interventional Pilot Study. *J. Clin. Med.* **2025**, *14*, 2083. https://doi.org/10.3390/jcm14062083.
16. Hudgens-Haney, M.E.; Clementz, B.A.; Ivleva, E.I.; Keshavan, M.S.; Pearlson, G.D.; Gershon, E.S.; Keedy, S.K.; Sweeney, J.A.; Gaudoux, F.; Bunouf, P.; et al. Cognitive Impairment and Diminished Neural Responses Constitute a Biomarker Signature of Negative Symptoms in Psychosis. *Schizophr. Bull.* **2020**, *46*, 1269–1281. https://doi.org/10.1093/schbul/sbaa001.
17. Van Gool, R.; Cay, M.; Van Der Heijden, H.; Golden, E.; Cao, A.; Ren, B.; Gonzalez-Heydrich, J.; Shinn, A.K.; Upadhyay, J. Neural and Behavioral Markers of Negative Symptoms and Cognitive Impairment in Patients with Psychosis. *J. Neurosci.* **2025**, *45*, e0039252025. https://doi.org/10.1523/JNEUROSCI.0039-25.2025.
18. Galderisi, S.; Mucci, A.; Dollfus, S.; Nordentoft, M.; Falkai, P.; Kaiser, S.; Giordano, G.M.; Vandevelde, A.; Nielsen, M.Ø.; Glenthøj, L.B.; et al. EPA Guidance on Assessment of Negative Symptoms in Schizophrenia. *Eur. Psychiatr.* **2021**, *64*, e23. https://doi.org/10.1192/j.eurpsy.2021.11.
19. Mucci, A.; Leucht, S.; Giordano, G.M.; Giuliani, L.; Wehr, S.; Weigel, L.; Galderisi, S. Assessment of Negative Symptoms in Schizophrenia: From the Consensus Conference-Derived Scales to Remote Digital Phenotyping. *Brain Sci.* **2025**, *15*, 83. https://doi.org/10.3390/brainsci15010083.
20. Demjaha, A.; Stahl, D.; Arango, C.; Glenthøj, B.; Rodriguez-Jimenez, R.; Díaz-Caneja, C.M.; Baandrup, L.; Ebdrup, B.; Paz Garcia-Portlla, M.; Diaz-Marsa, M.; et al. Confirmatory Factor Analysis of Competing PANSS Negative Symptom Models: Data from OPTiMiSE First-Episode Schizophrenia Study. *BJPsych Open* **2026**, *12*, e12. https://doi.org/10.1192/bjo.2025.10922.
21. Wallwork, R.S.; Fortgang, R.; Hashimoto, R.; Weinberger, D.R.; Dickinson, D. Searching for a Consensus Five-Factor Model of the Positive and Negative Syndrome Scale for Schizophrenia. *Schizophr. Res.* **2012**, *137*, 246–250. https://doi.org/10.1016/j.schres.2012.01.031.
22. Dal Santo, F.; García-Portilla, M.P.; Fernández-Egea, E.; González-Blanco, L.; Sáiz, P.A.; Giordano, G.M.; Galderisi, S.; Bobes, J. The Dimensional Structure of the Positive and Negative Syndrome Scale in First-Episode Schizophrenia Spectrum Disorders: An Exploratory Graph Analysis from the OPTiMiSE Trial. *Schizophrenia* **2024**, *10*, 81. https://doi.org/10.1038/s41537-024-00499-5.
23. Avan, M.; Hamurcu, H.D.; Pekiyi Avan, B.; Çayköylü, A. Association of Metabolic Syndrome with Comorbid Major Depressive Disorder in Remitted Schizophrenia: A Cross-Sectional Study. *BMC Psychiatry* **2026**, *26*, 391. https://doi.org/10.1186/s12888-026-08050-9.
24. Gao, S.; Xu, Q.; Han, Y.; Jiang, J.; Wu, F.; Peng, T.; Ling, C.; Ni, S.; Zhang, R.; Ming, Y.; et al. Relationship between Cognitive Impairments and Psychopathological Symptoms in Female Schizophrenia Subsequent to 8 Weeks Treatment with Antipsychotic Drugs. *BMC Psychiatry* **2025**, *25*, 211. https://doi.org/10.1186/s12888-025-06605-w.
25. Bellitti, M.; Lauber, M.V.; Searls, E.; Lin, H.; Au, R.; Kolachalama, V.B. Practice Effects on Digital Cognitive Assessment Tools: Insights from the Defense Automated Neurobehavioral Assessment Battery. *Alzheimer's Dement.* **2025**, *21*, e70644. https://doi.org/10.1002/alz.70644.
26. Winter, L.D.; Jelsma, A.; Vermeulen, J.M.; Tricht, M.V.; Weeghel, J.V.; Hasson-Ohayon, I.; Mulder, C.L.; Boonstra, N.; Veling, W.; Haan, L.D. Long-Term Changes in Cognition Among Patients with Schizophrenia Spectrum Disorders and Different Durations of Illness: A Meta-Analysis. *J. Clin. Psychiatry* **2024**, *85*, 23r15134. https://doi.org/10.4088/JCP.23r15134.
27. Goldberg, T.E.; Keefe, R.S.E.; Goldman, R.S.; Robinson, D.G.; Harvey, P.D. Circumstances Under Which Practice Does Not Make Perfect: A Review of the Practice Effect Literature in Schizophrenia and Its Relevance to Clinical Treatment Studies. *Neuropsychopharmacology* **2010**, *35*, 1053–1062. https://doi.org/10.1038/npp.2009.211.
28. Scharfen, J.; Peters, J.M.; Holling, H. Retest effects in cognitive ability tests: A meta-analysis. *Intelligence* **2018**, *67*, 44–66. https://doi.org/10.1016/j.intell.2018.01.003.
29. Allott, K.; Yassin, W.; Alameda, L.; Billah, T.; Borders, O.; Buccilli, K.; Carrión, R.E.; Castillo-Passi, R.I.; Cho, K.I.K.; Chin, K.; et al. Cognitive Assessment in the Accelerating Medicines Partnership® Schizophrenia Program: Harmonization Priorities and Strategies in a Diverse International Sample. *Schizophrenia* **2025**, *11*, 49. https://doi.org/10.1038/s41537-025-00578-1.

30. Fischer, R.; Nagel, M.; Schöttle, D.; Lüdecke, D.; Lassay, F.; Moritz, S.; Scheunemann, J. Metacognitive Training in the Acute Psychiatric Care Setting: Feasibility, Acceptability, and Safety. *Front. Psychol.* **2023**, *14*, 1247725. https://doi.org/10.3389/fpsyg.2023.1247725.
31. Lindgren, M.; Therman, S.; Avellan, A.; From, T.; Hietala, J.; Holm, M.; Ilonen, T.; Kieseppä, T.; Laurikainen, H.; Salokangas, R.K.R.; et al. Extrapyramidal Symptoms Predict Cognitive Performance after First-Episode Psychosis. *Schizophrenia* **2022**, *8*, 64. https://doi.org/10.1038/s41537-022-00270-8.
32. Gifford, G.; Cullen, A.E.; Vieira, S.; Searle, A.; McCutcheon, R.A.; Modinos, G.; Stone, W.S.; Hird, E.; Barnett, J.; Van Hell, H.H.; et al. PsyCog: A Computerised Mini Battery for Assessing Cognition in Psychosis. *Schizophr. Res. Cogn.* **2024**, *37*, 100310. https://doi.org/10.1016/j.scog.2024.100310.
33. Zorowitz, S.; Niv, Y. Improving the Reliability of Cognitive Task Measures: A Narrative Review. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **2023**, *8*, 789–797. https://doi.org/10.1016/j.bpsc.2023.02.004.
34. Morozova, A.; Zorkina, Y.; Abramova, O.; Pavlova, O.; Pavlov, K.; Soloveva, K.; Volkova, M.; Alekseeva, P.; Andryshchenko, A.; Kostyuk, G.; et al. Neurobiological Highlights of Cognitive Impairment in Psychiatric Disorders. *Int. J. Mol. Sci.* **2022**, *23*, 1217. https://doi.org/10.3390/ijms23031217.
35. Howes, O.; Marcinkowska, J.; Turkheimer, F.E.; Carr, R. Synaptic Changes in Psychiatric and Neurological Disorders: State-of-the Art of In Vivo Imaging. *Neuropsychopharmacology* **2025**, *50*, 164–183. https://doi.org/10.1038/s41386-024-01943-x.
36. Tudor, A.; Vasile, A.; Trifu, S.; Cristea, M. Morphological Classification and Changes in Dementia (Review). *Exp. Ther. Med.* **2021**, *23*, 33. https://doi.org/10.3892/etm.2021.10955.
37. Akl, E.; Dyrba, M.; Görß, D.; Schumacher, J.; Weber, M.-A. MRI for diagnosing dementia—Update 2025. *Rofo* **2025**, *197*, 1378–1387. https://doi.org/10.1055/a-2563-0725.
38. Benton, A.L.; Benton-Sivan, A.; Livinti, R.; Iliescu, D.; Mindu, K. *Benton Visual Retention Test (BVRT): Manual Tehnic si Interpretativ*, 5th ed.; Romanian Adaptation; TestCentral (D&D Consultants Grup): Bucharest, Romania, 2009.
39. Manna, C.B.G.; Filangieri, C.M.; Borod, J.C.; Alterescu, K.; Bender, H.A. Benton Visual Retention Test. In *Encyclopedia of Clinical Neuropsychology*; Kreutzer, J.S., DeLuca, J., Caplan, B., Eds.; Springer: Cham, Switzerland, 2018.
40. Knížková, K.; Keřková, B.; Večeřová, M.; Šustová, P.; Jonáš, J.; Siroňová, A.; Hrubý, A.; Rodriguez, M. Longitudinal Course of Core Cognitive Domains in First-Episode Acute and Transient Psychotic Disorders Compared with Schizophrenia. *Schizophr. Res. Cogn.* **2024**, *37*, 100311. https://doi.org/10.1016/j.scog.2024.100311.
41. Provenzani, U.; Salazar De Pablo, G.; Arribas, M.; Pillmann, F.; Fusar-Poli, P. Clinical Outcomes in Brief Psychotic Episodes: A Systematic Review and Meta-Analysis. *Epidemiol. Psychiatr. Sci.* **2021**, *30*, e71. https://doi.org/10.1017/S2045796021000548.
42. He, Q.; Wang, Y. Diagnostic Stability and Long-Term Outcomes of Acute and Transient Psychotic Disorders: A Systematic Review and Meta-Analysis. *Front. Psychiatry* **2026**, *17*, 1839599. https://doi.org/10.3389/fpsyt.2026.1839599.
43. World Health Organization. The ICD 10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines; World Health Organization: Geneva, Switzerland, 1992.
44. World Health Organization. *Clinical Descriptions and Diagnostic Requirements for ICD-11 Mental, Behavioural and Neurodevelopmental Disorders*, 1st ed.; World Health Organization: Geneva, Switzerland, 2024; ISBN 978-92-4-007726-3.
45. Elm, E.V.; Altman, D.G.; Egger, M.; Pocock, S.J.; Gøtzsche, P.C.; Vandenbroucke, J.P. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *BMJ* **2007**, *335*, 806–808. https://doi.org/10.1136/bmj.39335.541782.AD. Correction in *BMJ* **2007**, *335*, 0-a. https://doi.org/10.1136/bmj.39386.490150.94.
46. Murthy, S.; Marchello, C.S.; Hagedoorn, N.N.; Faigan, S.; Andrews, J.R.; Antillón, M.; Armah, G.E.; Bentsi-Enchill, A.D.; Birkhold, M.; Breiman, R.F.; et al. Strengthening the Reporting of Observational Studies in Epidemiology Enhanced Prevalence and Incidence Criteria (STROBE EPIC): An Extension of the STROBE Statement. *Ann. Intern. Med.* **2026**, *179*, 1280–1289. https://doi.org/10.7326/ANNALS-25-02412.
47. Kay, S.R.; Fiszbein, A.; Opler, L.A. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophr. Bull.* **1987**, *13*, 261–276. https://doi.org/10.1093/schbul/13.2.261.
48. Geck, S.; Roithmeier, M.; Bühner, M.; Wehr, S.; Weigel, L.; Priller, J.; Davis, J.M.; Leucht, S. COSMIN Systematic Review and Meta-Analysis of the Measurement Properties of the Positive and Negative Syndrome Scale (PANSS). *eClinicalMedicine* **2025**, *82*, 103155. https://doi.org/10.1016/j.eclinm.2025.103155.
49. Panss01—Data Structure—NDA. Available online: https://nda.nih.gov/data-structure/panss01 (accessed on 13 November 2026).
50. Student. The Probable Error of a Mean. *Biometrika* **1908**, *6*, 1. https://doi.org/10.2307/2331554.

51. Langenberg, B.; Janczyk, M.; Koob, V.; Kliegl, R.; Mayer, A. A Tutorial on Using the Paired t Test for Power Calculations in Repeated Measures ANOVA with Interactions. *Behav. Res.* **2022**, *55*, 2467–2484. https://doi.org/10.3758/s13428-022-01902-8.
52. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates, Publishers: Hillsdale, NJ, USA, 1988; ISBN 978-0-203-77158-7.
53. Lakens, D. Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs. *Front. Psychol.* **2013**, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863.
54. Giner-Sorolla, R.; Montoya, A.K.; Reifman, A.; Carpenter, T.; Lewis, N.A.; Aberson, C.L.; Bostyn, D.H.; Conrique, B.G.; Ng, B.W.; Schoemann, A.M.; et al. Power to Detect What? Considerations for Planning and Evaluating Sample Size. *Pers. Soc. Psychol. Rev.* **2024**, *28*, 276–301. https://doi.org/10.1177/10888683241228328.
55. Schubert, A.-L.; Steinhilber, M.; Kang, H.; Quintana, D.S. Improving Statistical Reporting in Psychology. *Commun. Psychol.* **2025**, *3*, 156. https://doi.org/10.1038/s44271-025-00356-w.
56. Chang, Y.-H.H.; Buras, M.R.; Davis, J.M.; Crowson, C.S. Avoiding Blunders When Analyzing Correlated Data, Clustered Data, or Repeated Measures. *J. Rheumatol.* **2023**, *50*, 1269–1272. https://doi.org/10.3899/jrheum.2022-1109.
57. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80. https://doi.org/10.2307/3001968.
58. Kim, J.; Kim, D.H.; Kwak, S.G. Comprehensive Guidelines for Appropriate Statistical Analysis Methods in Research. *Korean J Anesth.* **2024**, *77*, 503–517. https://doi.org/10.4097/kja.24016.
59. Pratt, J.W. Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures. *J. Am. Stat. Assoc.* **1959**, *54*, 655–667. https://doi.org/10.1080/01621459.1959.10501526.
60. Wicklin, R. Modifications of the Wilcoxon Signed Rank Test and Exact P-Values. *The DO Loop*, 24 July 2023.
61. Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904**, *15*, 72. https://doi.org/10.2307/1412159.
62. Geljic, M.; Mitchell, M.; Stevens, K.-A.; Holbrook, H.; Darke, H.; Goodbourn, P.; Damicoucas, C.; Asghari-Jafarabadi, M.; Sundram, S.; Carter, O. Visual Integration Deficits Associated with Psychosis Are Independent of Diagnosis. *Schizophrenia* **2025**, *11*, 58. https://doi.org/10.1038/s41537-025-00606-0.
63. Pearson, K. VII. Mathematical Contributions to the Theory of Evolution.—III. Regression, Heredity, and Panmixia. *Philos. Trans. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character* **1896**, 187, 253–318. https://doi.org/10.1098/rsta.1896.0007.
64. MacKinnon, J.G.; White, H. Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *J. Econom.* **1985**, *29*, 305–325. https://doi.org/10.1016/0304-4076(85)90158-7.
65. Rajh-Weber, H.; Huber, S.E.; Arendasy, M. A Practice-Oriented Guide to Statistical Inference in Linear Modeling for Non-Normal or Heteroskedastic Error Distributions. *Behav. Res.* **2025**, *57*, 338. https://doi.org/10.3758/s13428-025-02801-4.
66. Jones, L.; Barnett, A.; Vagenas, D. Common Misconceptions Held by Health Researchers When Interpreting Linear Regression Assumptions, a Cross-Sectional Study. *PLoS ONE* **2025**, *20*, e0299617. https://doi.org/10.1371/journal.pone.0299617.
67. Breusch, T.S.; Pagan, A.R. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* **1979**, *47*, 1287. https://doi.org/10.2307/1911963.
68. Marquardt, D.W. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics* **1970**, *12*, 591–612. https://doi.org/10.1080/00401706.1970.10488699.
69. Mullahy, J. Analyzing health outcomes measured as bounded counts. *J. Health Econ.* **2024**, *95*, 102875. https://doi.org/10.1016/j.jhealeco.2024.102875.
70. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1995**, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
71. Liang, K.-Y.; Zeger, S.L. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **1986**, *73*, 13–22. https://doi.org/10.1093/biomet/73.1.13.
72. Austin, P.C.; Kapral, M.K.; Vyas, M.V.; Fang, J.; Yu, A.Y.X. Using Multilevel Models and Generalized Estimating Equation Models to Account for Clustering in Neurology Clinical Research. *Neurology* **2024**, *103*, e209947. https://doi.org/10.1212/WNL.0000000000209947.
73. Zhang, Y.; Preisser, J.S. Finite-Sample Adjustments in Estimating Equations for Correlated Overdispersed Count Outcomes with Application to Cluster Randomized Trials. *Stat. Med.* **2026**, *45*, e70698. https://doi.org/10.1002/sim.70698.
74. Wald, A. Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large. *Trans. Amer. Math. Soc.* **1943**, *54*, 426–482. https://doi.org/10.1090/S0002-9947-1943-0012401-3.
75. Mancl, L.A.; DeRouen, T.A. A Covariance Estimator for GEE with Improved Small-Sample Properties. *Biometrics* **2001**, *57*, 126–134. https://doi.org/10.1111/j.0006-341X.2001.00126.x.

76. Hemming, K.; Thompson, J.; Kristunas, C.; Forbes, A.; Watson, S.; Bowden, R.; McKenzie, J.E.; Kasza, J. The Performance of Small Sample Correction Methods for Controlling Type I Error When Analyzing Parallel Cluster Randomized Trials: A Systematic Review of Simulation Studies. *J. Clin. Epidemiol.* **2025**, *185*, 111838. https://doi.org/10.1016/j.jclinepi.2025.111838.
77. Welch, B.L. The Significance of the Difference Between Two Means When the Population Variances Are Unequal. *Biometrika* **1938**, *29*, 350–362. https://doi.org/10.1093/biomet/29.3-4.350.
78. Curtis, D. Welch's t Test Is More Sensitive to Real World Violations of Distributional Assumptions than Student's t Test but Logistic Regression Is More Robust than Either. *Stat. Pap.* **2024**, *65*, 3981–3989. https://doi.org/10.1007/s00362-024-01531-7.
79. Hoenig, J.M.; Heisey, D.M. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *Am. Stat.* **2001**, *55*, 19–24. https://doi.org/10.1198/000313001300339897.
80. Berg, R.M.G.; Christensen, R.; Ried-Larsen, M. The Corruption of Power: On the Use and Abuse of a Pre-trial Concept. *Exp. Physiol.* **2024**, *109*, 317–319. https://doi.org/10.1113/EP091560.
81. Python Software Foundation. *Python Language Reference*, version 3.12.14; Python Software Foundation: Beaverton, OR, USA, 2026. Available online: https://www.python.org/downloads/release/python-31214/ (accessed on 2 October 2026).
82. McKinney, W. *Data Structures for Statistical Computing in Python*; SCIPY: Austin, TX, USA, 2010; pp. 56–61.
83. The Pandas Development Team. *Pandas-Dev/Pandas: Pandas*; Zenodo: Geneva, Switzerland, 2024.
84. Harris, C.R.; Millman, K.J.; Van Der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array Programming with NumPy. *Nature* **2020**, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.
85. NumPy 2.5.4 Release Notes—NumPy v2.6.Dev0 Manual. Available online: https://numpy.org/devdocs/release/2.5.4-notes.html (accessed on 13 November 2026).
86. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2. Correction in *Nat. Methods* **2020**, *17*, 352. https://doi.org/10.1038/s41592-020-0772-5.
87. Gommers, R.; Virtanen, P.; Haberland, M.; Burovski, E.; Reddy, T.; Weckesser, W.; Oliphant, T.E.; Nelson, A.; Cournapeau, D.; Roy, P.; et al. *Scipy/Scipy: SciPy 1.16.3*; Zenodo: Geneva, Switzerland, 2025.
88. Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python.; Austin, Texas, 2010; pp. 92–96.
89. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.
90. The Matplotlib Development Team. *Matplotlib: Visualization with Python*, version v3.11.2; Zenodo: Geneva, Switzerland, 2026.
91. NumPy Developers Random Generator. In *NumPy Reference Manual*; NumPy Developers: Beaverton, OR, USA, 2025.
92. Calamia, M.; Markon, K.; Tranel, D. Scoring Higher the Second Time Around: Meta-Analyses of Practice Effects in Neuropsychological Assessment. *Clin. Neuropsychol.* **2012**, *26*, 543–570. https://doi.org/10.1080/13854046.2012.680913.
93. Goldberg, T.E.; Goldman, R.S.; Burdick, K.E.; Malhotra, A.K.; Lencz, T.; Patel, R.C.; Woerner, M.G.; Schooler, N.R.; Kane, J.M.; Robinson, D.G. Cognitive Improvement After Treatment with Second-Generation Antipsychotic Medications in First-Episode Schizophrenia: Is It a Practice Effect? *Arch. Gen. Psychiatry* **2007**, *64*, 1115. https://doi.org/10.1001/archpsyc.64.10.1115.
94. Demirlek, C.; Keshavan, M.S.; Murray, R.M.; Bora, E. No Evidence for Cognitive Decline in First-Episode Psychosis: A Meta-Analysis of Long-Term Longitudinal Studies. *Schizophr. Bull.* **2026**, *52*, sbaf237. https://doi.org/10.1093/schbul/sbaf237.
95. Keefe, R.S.E.; Davis, V.G.; Harvey, P.D.; Atkins, A.S.; Haig, G.M.; Hagino, O.; Marder, S.; Hilt, D.C.; Umbricht, D. Placebo Response and Practice Effects in Schizophrenia Cognition Trials. *JAMA Psychiatry* **2017**, *74*, 807. https://doi.org/10.1001/jamapsychiatry.2017.1574.
96. Horan, W.P.; Kalali, A.; Brannan, S.K.; Drevets, W.; Leoni, M.; Mahableshwarkar, A.; Martin, W.J.; Rao, S.; Reuteman-Fowler, C.; Sauder, C.; et al. Towards Enhancing Drug Development Methodology to Treat Cognitive Impairment Associated with Schizophrenia and Other Neuropsychiatric Conditions: Insights from 2 Decades of Clinical Trials. *Schizophr. Bull.* **2025**, *51*, 262–273. https://doi.org/10.1093/schbul/sbae151.
97. Duff, K. Evidence-Based Indicators of Neuropsychological Change in the Individual Patient: Relevant Concepts and Methods. *Arch. Clin. Neuropsychol.* **2012**, *27*, 248–261. https://doi.org/10.1093/arclin/acr120.
98. Vita, A.; Gaebel, W.; Mucci, A.; Sachs, G.; Erfurth, A.; Barlati, S.; Zanca, F.; Giordano, G.M.; Birkedal Glenthøj, L.; Nordentoft, M.; et al. European Psychiatric Association Guidance on Assessment of Cognitive Impairment in Schizophrenia. *Eur. Psychiatr.* **2022**, *65*, e58. https://doi.org/10.1192/j.eurpsy.2022.2316.
99. Rek-Owodziń, K.; Tyburski, E.; Plichta, P.; Waszczuk, K.; Bielecki, M.; Wietrzyński, K.; Podwalski, P.; Rudkowski, K.; Michalczyk, A.; Grążlewski, T.; et al. The Relationship between Cognitive Functions and Psychopathological Symptoms in First Episode Psychosis and Chronic Schizophrenia. *J. Clin. Med.* **2022**, *11*, 2619. https://doi.org/10.3390/jcm11092619.

100. Dukes, O.; Shahn, Z.; Renson, A. Change Scores and Baseline Adjustment: Splitting the Difference (in Differences). *Int. J. Epidemiol.* **2025**, *54*, dyaf042. https://doi.org/10.1093/ije/dyaf042.
101. Loh, W.W.; Ren, D. Adjusting for Baseline Measurements of the Mediators and Outcome as a First Step Toward Eliminating Confounding Biases in Mediation Analysis. *Perspect. Psychol. Sci.* **2023**, *18*, 1254–1266. https://doi.org/10.1177/17456916221134573.
102. Fett, A.-K.J.; Velthorst, E.; Reichenberg, A.; Ruggero, C.J.; Callahan, J.L.; Fochtmann, L.J.; Carlson, G.A.; Perlman, G.; Bromet, E.J.; Kotov, R. Long-term Changes in Cognitive Functioning in Individuals With Psychotic Disorders: Findings From the Suffolk County Mental Health Project. *JAMA Psychiatry* **2020**, *77*, 387–396. https://doi.org/10.1001/jamapsychiatry.2019.3993.
103. Arnautovska, U.; Vitangcol, K.; Kesby, J.P.; Warren, N.; Rossell, S.L.; Neill, E.; Harris, A.; Galletly, C.; Castle, D.; Siskind, D. Verbal and Visual Learning Ability in Patients with Treatment-Resistant Schizophrenia: A 1-Year Follow-up Study. *Schizophr. Res. Cogn.* **2023**, *33*, 100283. https://doi.org/10.1016/j.scog.2023.100283.
104. Kalisova, L.; Michalec, J.; Dechterenko, F.; Silhan, P.; Hyza, M.; Chlebovcova, M.; Brenova, M.; Bezdicek, O. Impact of Cognitive Performance and Negative Symptoms on Psychosocial Functioning in Czech Schizophrenia Patients. *Schizophrenia* **2023**, *9*, 43. https://doi.org/10.1038/s41537-023-00374-9.
105. Blotner, J.; Hechinger, R.; Shenton, M.E.; Coleman, M.J.; Öngür, D.; Lewandowski, K.E. Negative symptoms, cognition, and community functioning in early psychosis. *Schizophr. Res.* **2025**, *281*, 260–267. https://doi.org/10.1016/j.schres.2025.05.015.
106. Seel, N.M. Experimental and Quasi-Experimental Designs for Research on Learning. In *Encyclopedia of the Sciences of Learning*; Seel, N.M., Ed.; Springer: Boston, MA, USA, 2012; pp. 1223–1229.
107. Karr, J.E.; Ingram, E.O.; Pinheiro, C.N.; Ali, S.; Iverson, G.L. Test–Retest Reliability and Reliable Change on the NIH Toolbox Cognition Battery. *Arch. Clin. Neuropsychol.* **2024**, *39*, 702–713. https://doi.org/10.1093/arclin/acae011.
108. Dahabreh, I.J.; Bibbins-Domingo, K. Causal Inference About the Effects of Interventions from Observational Studies in Medical Journals. *JAMA* **2024**, *331*, 1845. https://doi.org/10.1001/jama.2024.7741.
109. Jacobson, N.S.; Truax, P. Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research. *J. Consult. Clin. Psychol.* **1991**, *59*, 12–19. https://doi.org/10.1037/0022-006X.59.1.12.
110. Nuechterlein, K.H.; Nasrallah, H.; Velligan, D. Measuring Cognitive Impairments Associated with Schizophrenia in Clinical Practice: Overview of Current Challenges and Future Opportunities. *Schizophr. Bull.* **2025**, *51*, 401–421. https://doi.org/10.1093/schbul/sbae051.
111. Oldham, P.D. A Note on the Analysis of Repeated Measurements of the Same Subjects. *J. Chronic Dis.* **1962**, *15*, 969–977. https://doi.org/10.1016/0021-9681(62)90116-9.
112. Barnett, A.G. Regression to the Mean: What It Is and How to Deal with It. *Int. J. Epidemiol.* **2004**, *34*, 215–220. https://doi.org/10.1093/ije/dyh299.
113. Ratcliffe, L.N.; Hale, A.C.; McDonald, T.; Hewitt, K.C.; Nguyen, C.M.; Spencer, R.J.; Loring, D.W. The Montreal Cognitive Assessment: Norms and Reliable Change Indices for Standard and MoCA-22 Administrations. *Arch. Clin. Neuropsychol.* **2024**, *39*, 747–765. https://doi.org/10.1093/arclin/acae013.
114. Papke, L.; Wooldridge, J. *Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates*; National Bureau of Economic Research: Cambridge, MA, USA, 1993; p. t0147.
115. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. https://doi.org/10.1177/001316446002000104.
116. English, J.; Meyers, J.E.; Gran, J.M.; Chatmon, L.M.; Gomez, B.M.; Kockler, T.; Sica, R.M.; Greco, S.P.; Shah, M.P. Advancing Diagnostic Consistency: Statistical Methods and Interrater Reliability in Neuropsychological Assessments with the MNB. *Appl. Neuropsychol. Adult* **2026**, *33*, 1132–1136. https://doi.org/10.1080/23279095.2024.2443954.
117. Simmons, J.P.; Nelson, L.D.; Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.* **2011**, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632.
118. Fingerhut, A.; Uranues, S.; Dziri, C.; Ma, J.; Vernerey, D.; Kurihara, H.; Stiegler, P. Interaction Analysis of Subgroup Effects in Randomized Trials: The Essential Methodological Points. *Sci. Rep.* **2024**, *14*, 12619. https://doi.org/10.1038/s41598-024-62896-1.
119. Berkson, J. Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biom. Bull.* **1946**, *2*, 47. https://doi.org/10.2307/3002000.

120. Yoon, J.; Mayer, M.R.; Berro, T.; Brazis, S.; Kantrowitz, J.T. Knowing Is Half the Battle: The Factors Leading to Efficient Recruitment of Representative Samples in Schizophrenia Research. *Pharm. Med.* **2025**, *39*, 29–38. https://doi.org/10.1007/s40290-024-00545-8.
121. Bross, I. Misclassification in 2 X 2 Tables. *Biometrics* **1954**, *10*, 478. https://doi.org/10.2307/3001619.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.