

Stochastic Competence of In-Service Mathematics Teachers in Uzbekistan: Comparing Self-Assessment with Objective Evidence Under the National Curriculum

Tashtemirova Nargiza Nematovna

Gulistan State Pedagogical Institute, Gulistan, Uzbekistan

Received: 04 August 2026 Accepted: 31 August 2026 Published: 29 September 2026

ABSTRACT

Probability and statistics now form a separate content area of Uzbekistan's National Curriculum, which places interpretive and pedagogical demands on mathematics teachers. Empirical evidence on in-service teachers' stochastic competence in Uzbekistan remains limited, and continuous professional development (CPD) needs are commonly identified through self-assessment alone. This study compared self-assessed and objectively assessed stochastic competence across four curriculum-aligned components: probability–statistical thinking, data analysis and interpretation, statistical modelling, and professional-reflexive activity. A convergent mixed-methods design combined a self-assessment survey with open-ended questions ($N = 175$; Cronbach's $\alpha = 0.74–0.81$) and an objective competence test administered to a separate, unlinked sample ($N = 242$; KR-20 = 0.78 for the closed items). Because the samples were independent, the two sources were compared at the group and domain level only. The mean overall self-rating was 3.49 out of 5, and prior stochastic professional development was associated with higher self-assessed competence (Mann–Whitney $U = 4253.0$, $p = 0.005$, $r = 0.21$; sensitivity analysis: $t = 3.20$, $p = 0.002$, $d = 0.52$). The mean objective competence index was 45.5% ($Md = 44.4$, $SD = 22.3$); 58.7% of teachers scored below 50%, 26.4% between 50% and 69%, and 14.9% at 70% or above. Component indices were 46.5% for probability–statistical thinking, 40.3% for data analysis and interpretation, and 50.2% for statistical modelling, with a derived estimate of 45.0% for professional-reflexive activity. Correct-response rates were 22.3% on a base-rate task, 26.9% on a truncated-axis task, and 27.3% on a correlation–causation task. The two sources showed partial convergence in conditional probabilistic reasoning but divergence in data analysis and interpretation. The findings suggest that self-assessment alone may not adequately identify CPD needs and that practice-based CPD should target data interpretation and conditional probabilistic reasoning.

Keywords: Stochastic competence; statistical knowledge for teaching; self-assessment; objective competence assessment; in-service mathematics teachers; National Curriculum; Uzbekistan.

INTRODUCTION

1. Background and research gap

The ability to interpret statistical information critically, to reason under uncertainty, and to draw conclusions from data has become part of basic literacy in data-rich societies (Gal, 2002; Wild & Pfannkuch, 1999). School mathematics has responded by treating data, statistics, and probability

as an interconnected stochastic strand aimed at pupils' statistical literacy, reasoning, and thinking (Ben-Zvi & Garfield, 2004; Batanero et al., 2011). These outcomes depend on teachers. Teaching stochastic content requires more than procedural fluency: teachers need to interpret variability and context, anticipate pupils' misconceptions, and select tasks and representations that make reasoning under uncertainty accessible (Batanero et al., 2011; Groth,

2007).

International research indicates that teachers themselves often experience difficulties with intuitive interpretations of probability, variability, sampling and inference, and the distinction between correlation and causation (Batanero et al., 2011; Batanero & Chernoff, 2018; Batanero & Álvarez-Arroyo, 2024). These difficulties matter because they concern precisely the ideas that teachers are expected to develop, and to diagnose, in their pupils.

In Uzbekistan, elements of combinatorics, probability, and statistics entered school mathematics with the competency-based state educational standard of 2017. The National Curriculum, introduced under the national programme for the development of school education in 2022–2026 (President of the Republic of Uzbekistan, 2022), systematised "Probability and Statistics" as a separate content area with four strands: representing and interpreting data (E1), statistical processing of data (E2), combinatorics (E3), and probability (E4). Its learning objectives extend beyond computation to decisions and forecasts based on real data, the evaluation of sampling methods, and the interpretation of uncertainty in real-life and environmental contexts. As broader context, Uzbekistan's results in PISA 2022 (364 points in mathematics) and TIMSS 2023 (443 points in Grade 4 and 421 in Grade 8) underline the need to strengthen mathematics teaching (OECD, 2023; von Davier et al., 2024). These studies measure student achievement, not teachers' stochastic competence, and are cited here only as contextual evidence.

In 2024, a presidential resolution introduced school-based forms of continuous professional development (CPD) for teachers (President of the Republic of Uzbekistan, 2024). Designing CPD content for the new curriculum requires evidence on where teachers' needs lie. Such needs are often identified through self-assessment, yet self-assessments are known to align only imperfectly with measured performance (Dunning et al., 2004). Objective diagnostics, conversely, do not capture teachers' perceived needs or preferred forms of support. Combining the two sources can therefore provide a more complete basis for CPD planning, provided that the comparison respects the structure of the data.

In the CIS tradition, the methodological preparation of mathematics teachers for teaching stochastics has been studied for some time (Selyutin, 2001), whereas research

on stochastics education in Uzbekistan has focused mainly on pupils (Xonqulov, 2018). Consequently, empirical evidence on in-service mathematics teachers' stochastic competence in Uzbekistan remains limited, particularly in relation to the explicit demands of the new National Curriculum.

2. Theoretical framework

The study draws on three nested models of teacher knowledge. Shulman's (1986) pedagogical content knowledge (PCK) describes the knowledge that makes subject matter comprehensible to learners, including suitable representations and awareness of learners' conceptions. Ball et al. (2008) refined PCK for mathematics as mathematical knowledge for teaching (MKT), specifying the knowledge needed to analyse pupils' solutions, locate the sources of errors, and choose representations. Groth (2007) argued that statistics requires a further, domain-specific extension, statistical knowledge for teaching (SKT), because statistical reasoning depends on the data context, variability, and the limits of empirical inference, which a general MKT model does not fully capture. PCK thus provides the general frame, MKT its mathematical specification, and SKT the stochastic specification adopted here.

On this basis, a mathematics teacher's stochastic competence is defined as the integration of (a) conceptual and content knowledge of data, statistics, combinatorics, and probability; (b) probabilistic and statistical reasoning under variability and uncertainty; (c) pedagogical knowledge for teaching this content; and (d) professional-reflexive competence, that is, the capacity to analyse pupils' reasoning and one's own teaching decisions. The construct is operationalised in four components. Probability–statistical thinking (I) refers to reasoning about random events and uncertainty and draws on research on probabilistic reasoning (Batanero & Chernoff, 2018). Data analysis and interpretation (II) refers to the critical reading of data and displays in context, as described in models of statistical literacy (Gal, 2002). Statistical modelling (III) refers to representing real situations with statistical models and justifying inferences and forecasts, in line with the investigative cycle of statistical thinking (Wild & Pfannkuch, 1999). Professional-reflexive activity (IV) refers to the pedagogical dimension shared by PCK, MKT, and SKT: diagnosing pupils' reasoning and justifying teaching decisions. Table 1 presents the operational definitions, curriculum alignment, and assessment

evidence for each component.

Table 1. Stochastic competence components: operational definitions, National Curriculum alignment, and assessment evidence

Competence component	Operational definition	Primary NC strand(s)	Assessment evidence
I. Probability–statistical thinking	Modelling combinatorial situations; quantifying and interpreting random events, including conditional probability and base rates; recognising typical probabilistic misconceptions	E3, E4	Survey: combinatorics and probability blocks; test: Block I (closed items)
II. Data analysis and interpretation	Reading, comparing, and critically appraising data displays; interpreting data in context; distinguishing correlation from causation	E1	Survey: statistics block (partly); test: Block II (closed items)
III. Statistical modelling	Using measures of central tendency and dispersion, evaluating samples, and justifying inferences and forecasts from data	E2	Survey: statistics block (partly); test: Block III (closed items)
IV. Professional-reflexive activity	Explaining and diagnosing pupils' reasoning, justifying pedagogical decisions, designing tasks, and using representations and digital tools	E1–E4	Survey: teaching-skill block and open-ended questions; test: Block IV (open-response pedagogical tasks)

Note. The alignment indicates primary correspondence rather than a one-to-one mapping: Component I draws on two strands, Component IV spans all strands, and the survey statistics block spans Components II and III. Survey blocks were organised by content topic and therefore correspond to the components only approximately.

3. Purpose, research questions, and contributions

The purpose of the study was to compare in-service mathematics teachers' self-assessed and objectively assessed stochastic competence at the group and domain level across the four curriculum-aligned components, and to identify reasoning difficulties and CPD needs that could inform professional development. Three research questions were addressed:

RQ1. How do in-service mathematics teachers assess their stochastic content knowledge and teaching skill, and are their self-assessments associated with prior stochastic professional development, teaching experience, and region?

RQ2. What level of stochastic competence do teachers demonstrate on an objective test, and which reasoning tasks are most difficult?

RQ3. Where do self-assessment and objective evidence converge or diverge at the domain level, and what CPD needs do teachers report?

The study makes three contributions. First, it operationalises stochastic competence for in-service mathematics teachers in relation to the National Curriculum. Second, it combines self-assessment and objective evidence while explicitly limiting comparisons to the group and domain level, because the two samples are independent. Third, it identifies specific probabilistic and statistical reasoning difficulties that can inform evidence-based professional development.

METHODS

1. Research design

The study followed a convergent mixed-methods design.

The quantitative strands were a self-assessment survey and an objective competence test. The qualitative strands were the open-ended survey responses, analysed thematically, and the open-response pedagogical tasks of the test, scored against an analytic rubric. The strands were collected separately, analysed independently, and integrated at the interpretation stage through a joint display organised by competence component, in which each domain was classified as showing convergence, partial convergence, or divergence (Fetters et al., 2013).

The survey and the test were administered to two independent, unlinked convenience samples. Consequently, no individual-level correlation, individual-level calibration, or individual confidence–competence discrepancy could be analysed. All comparisons between the two sources are group-level comparisons, and their results are reported as domain-level convergence or divergence.

2. Participants and sampling

The survey sample comprised 175 mathematics teachers from general secondary schools in Xorazm (77), Surxondaryo (60), and Sirdaryo (22) regions and Tashkent city (16). By teaching experience, 51 had 1–5 years, 33 had 6–10 years, 49 had 11–20 years, and 42 had more than 20 years. By qualification category, 55 held specialist status, 72 the second category, 30 the first category, and 18 the highest category. Overall, 118 reported prior professional development related to stochastic content and 57 reported none.

The objective test was administered in a separate diagnostic study and completed by 242 in-service mathematics teachers; background characteristics of this sample are not reported. Both samples were obtained through convenience sampling, and neither was designed to be nationally representative.

3. Instruments and scoring

Self-assessment survey. The electronic survey used a five-point self-rating scale (1 = very low, 5 = high) across four knowledge-and-skill blocks: combinatorics (4 items, $\alpha = 0.81$), probability (5 items, $\alpha = 0.79$), statistics (5 items, $\alpha = 0.78$), and teaching skill (5 items, $\alpha = 0.74$). A teaching-

barriers block (3 items) was analysed descriptively only. The survey also contained a holistic rating of overall stochastic preparedness, items on preferred CPD formats, and two open-ended questions on difficulties in teaching stochastic content and on CPD needs.

Objective competence test. The test comprised four blocks corresponding to Components I–IV, each carrying a maximum of 25 points, for a total maximum score of 100 (Table 2). Blocks I–III consisted of 18 single-answer multiple-choice items scored 1/0, which addressed conditional probability and base rates, typical probabilistic misconceptions, sampling, statistical inference, the reading of statistical displays, and the distinction between correlation and causation. Block IV consisted of open-response pedagogical tasks that required teachers to explain a solution or a faulty argument, justify a pedagogical decision, and propose a stochastic task; these tasks were scored by a single rater against an analytic rubric. Raw block scores were rescaled to the 25-point block maximum, and competence indices were computed as $CI = (\text{obtained score} / \text{maximum possible score}) \times 100\%$ for each component and for the test as a whole; for the whole test, the index therefore equals the total score out of 100.

Component indices for Blocks I–III were reported directly. Because the four blocks carried equal maximum scores (25 points each), the Component IV index was derived algebraically from the overall competence index and the reported indices for Components I–III: $4 \times 45.5 - (46.5 + 40.3 + 50.2) = 45.0\%$. Because the reported component indices were rounded, the derived estimate is subject to a rounding tolerance of approximately ± 0.3 percentage points. It is therefore reported as a derived Component IV estimate rather than a directly measured Component IV score. The number of open-response tasks and the rubric dimensions are not reported in this article.

Overall indices were grouped into descriptive performance bands ($<50\%$, $50\text{--}69\%$, $\geq 70\%$). Because no standard-setting study was conducted, the bands are not interpreted as validated competence levels. The instrument and bands are specific to this diagnostic study and differ from the criterion test and level thresholds used in the author's subsequent intervention research.

Table 2. Structure and scoring of the objective competence test

Block / component	Assessment format	Scoring	Maximum score (points)
I. Probability–statistical thinking	Single-answer multiple-choice items	Dichotomous (1/0), rescaled	25
II. Data analysis and interpretation	Single-answer multiple-choice items	Dichotomous (1/0), rescaled	25
III. Statistical modelling	Single-answer multiple-choice items	Dichotomous (1/0), rescaled	25
IV. Professional-reflexive activity	Open-response pedagogical tasks	Analytic rubric, single rater, rescaled	25
Total	18 closed items (Blocks I–III) and open-response tasks (Block IV)	Sum of block scores	100

Note. Each block contributes 25% of the total maximum score. The distribution of the 18 closed items across Blocks I–III and the number of open-response tasks are not reported.

4. Validity and psychometric evidence

Different forms of evidence were available for different parts of the instruments. For content and construct representation, experts reviewed the survey items against the competence components and the National Curriculum learning objectives, and three experts in stochastics, mathematics education, and educational measurement reviewed the test items for construct fit, coverage, clarity, and difficulty. For internal consistency, Cronbach's α ranged from 0.74 to 0.81 across the survey blocks, and KR-20 was 0.78 for the 18 closed test items. For item functioning, item difficulty ranged from $p = 0.22$ to 0.66 (mean 0.45), upper–lower 27% discrimination ranged from $D = 0.29$ to 0.77 (mean 0.54), and point-biserial correlations ranged from 0.26 to 0.49 ; no item was removed. The test was not piloted separately; its psychometric properties were estimated on the main sample. These indices support the internal consistency and item functioning of the closed items but do not by themselves establish full validity. The reliability of the open-response scores could not be estimated because a single rater scored them.

5. Data analysis

For RQ1, survey data were summarised with means, standard deviations, and percentage distributions, and a composite score was formed from the 19 knowledge-and-skill items. Normality of the composite was examined with the Shapiro–Wilk test and Q–Q plots, and homogeneity of variances with Levene's test. Because the composite

departed from normality in some groups, the Mann–Whitney U test (effect size r) was used to compare teachers with and without prior stochastic professional development, and the Kruskal–Wallis H test to compare the four experience groups; regional differences were also examined. Welch's t-test and one-way ANOVA were run as sensitivity analyses to check whether conclusions depended on the choice of test. Spearman's ρ was used for associations between survey blocks because it assumes only monotonic relations between ordinal-type scores. All tests were two-tailed at $\alpha = 0.05$.

For RQ2, the overall competence index, the descriptive performance bands, the component indices, and item-level correct-response rates were summarised descriptively.

For RQ3, open-ended responses were analysed inductively following Braun and Clarke's (2006) six-phase thematic analysis; a response could be coded under several themes. Survey indicators, test indices, and open-ended themes were then juxtaposed in a joint display (Fetters et al., 2013). A domain was classified as showing convergence when all sources pointed to the same strength or need, partial convergence when the sources agreed on the direction of the identified need but not on all indicators, and divergence when self-assessment indicated relative strength while objective evidence indicated relative weakness, or vice versa. Because the scales differ, these classifications concern the direction of evidence, not the equality of numerical values.

6. Ethics

Participation was voluntary, and informed consent was obtained before data collection. No directly identifying information entered the database, and results were analysed in aggregate only.

RESULTS

1. Self-assessed competence (RQ1)

The mean holistic rating of overall stochastic preparedness was 3.49. Of the respondents, 6.3% rated their preparation as low, 42.9% as average, 46.9% as good, and 4.0% as high. Block means ranged from 3.52 to 3.75, with probability (3.58) and teaching skill (3.52) lowest (Figure 1a). At item level, the lowest knowledge rating was for conditional probability ($M = 3.45$), followed by sampling and statistical forecasts ($M = 3.58$); mean, mode, and median received the highest rating ($M = 4.09$). In the teaching-skill block, modern methodological approaches ($M = 3.37$) and digital simulation and visualisation tools ($M = 3.48$) were rated lowest.

Teachers who reported prior stochastic professional development had higher composite self-assessment scores than those who did not ($U = 4253.0$, $p = 0.005$, $r = 0.21$); the parametric sensitivity analysis yielded the same result

($t = 3.20$, $p = 0.002$, $d = 0.52$). This association concerns self-assessed competence; it could not be examined for objective test performance. No statistically significant differences were detected across the four experience groups or across regions ($p > 0.05$); given the unequal regional group sizes, the regional result should be interpreted with caution. The four survey blocks were positively intercorrelated ($\rho = 0.60$ – 0.76 , all $p < 0.001$).

2. Objective competence (RQ2)

The mean overall competence index was 45.5% ($Md = 44.4$, $SD = 22.3$). In the descriptive performance bands, 58.7% of teachers scored below 50%, 26.4% between 50% and 69%, and 14.9% at 70% or above. Data analysis and interpretation had the lowest component index (40.3%), followed by probability–statistical thinking (46.5%) and statistical modelling (50.2%); the derived estimate for professional-reflexive activity was 45.0% (Figure 1b). The lowest item-level results concerned contextual reasoning: 22.3% of teachers correctly answered a medical-test task requiring the joint interpretation of conditional probability and base rates, 26.9% a task on how a truncated axis distorts a chart, and 27.3% a task distinguishing correlation from causation.

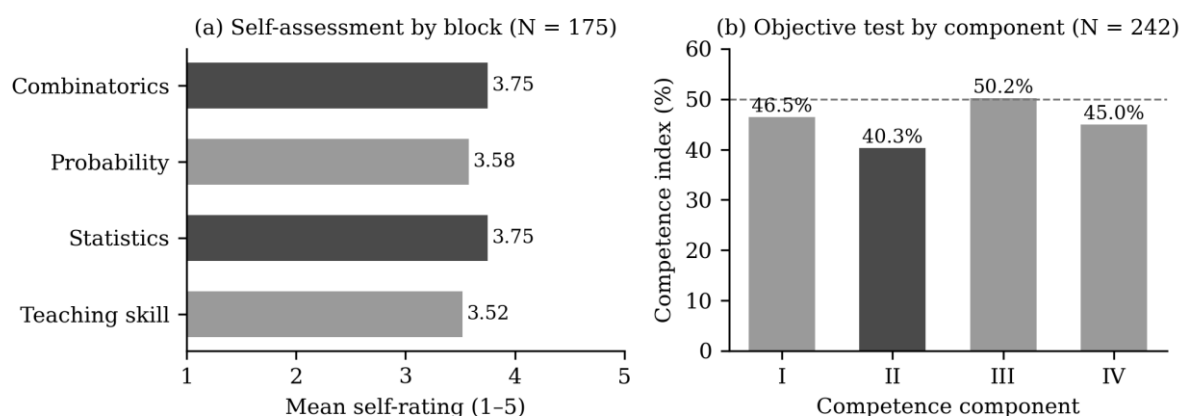


Figure 1. Mean self-assessment by survey block (a) and objective competence index by component (b). I = probability–statistical thinking; II = data analysis and interpretation; III = statistical modelling; IV = professional-reflexive activity. The value for Component IV is a derived estimate (Section 2.3).

3. Open-ended responses and domain-level integration (RQ3)

The open-ended questions yielded 84 substantive responses on improving the teaching of stochastics and 78

on improving CPD content. Five themes were identified: greater use of practical, real-life examples (34 coded responses), more time for stochastic content (19), strengthening the methodology of teaching stochastics (9), digital tools and distance formats (9), and improved textbooks and methodological resources (7). Practical

training and master classes were the most frequently preferred formats. Table 3 presents the joint display.

Table 3. Joint display of self-assessment, objective, and open-ended evidence by competence component

Component	Self-assessment (1–5)	Objective test	Open-ended responses	Domain-level status
I. Probability–statistical thinking	Probability block 3.58; conditional probability 3.45 (lowest knowledge item)	46.5%; base-rate task 22.3% (lowest task)	Deeper work on demanding topics	Partial convergence
II. Data analysis and interpretation	Sampling and forecasts 3.58; mean, mode, and median 4.09 (highest item)	40.3% (lowest component); truncated axis 26.9%; correlation–causation 27.3%	Practical, real-life examples (34)	Divergence
III. Statistical modelling	Statistics block 3.75 ^a	50.2% (highest component)	No component-specific theme	Partial convergence
IV. Professional-reflexive activity	Teaching skill 3.52 (lowest block); modern approaches 3.37; digital tools 3.48	45.0% (derived estimate) ^b	Methodology (9); digital tools and distance formats (9); practical training	Partial convergence in identified CPD needs

Note. The survey (N = 175) and test (N = 242) samples are independent; comparisons are descriptive and made at the group and domain level. Status categories concern the direction of the evidence, not the equality of values measured on different scales (Section 2.5). ^aThe statistics block spans Components II and III. ^bDerived from the overall index and the Component I–III indices under equal block maxima of 25 points (rounding tolerance ± 0.3 percentage points).

In Component I, the lowest-rated knowledge item and the lowest-scoring test task both concerned conditional probability. In Component II, descriptive statistics received the highest item-level self-rating, whereas the test component requiring the interpretation of data and displays produced the lowest index. In Component III, the highest self-rated block coincided with the highest objective index, which nevertheless remained at approximately half of the maximum score. In Component IV, the lowest-rated survey block, the open-ended requests for methodological and digital-tool support, and the derived test estimate below 50% indicated needs in the same direction.

DISCUSSION

1. Principal findings

The two independent samples reveal domain-level convergence and divergence between self-assessed and objectively assessed stochastic competence. At the group level, self-assessments were moderately favourable (3.49 out of 5), whereas objective performance was substantially lower, with most test-takers scoring below 50%. Because

the instruments use different scales and were administered to different teachers, this contrast does not indicate how any individual teacher's self-assessment relates to his or her performance; the substantive evidence lies in the domain-level pattern of convergence and divergence.

2. Domain-specific interpretation

Data analysis and interpretation. The divergence in this domain is the study's central finding. Descriptive procedures received the highest self-rating (mean, mode, and median: 4.09), yet the component requiring the interpretation of data and displays produced the lowest objective index (40.3%), and only about a quarter of teachers identified a truncated-axis distortion (26.9%) or distinguished correlation from causation (27.3%). These results may indicate that familiarity with descriptive procedures does not necessarily imply competence in the critical interpretation of data. This interpretive dimension, which depends on context, variability, and the limits of inference, is what distinguishes SKT from general mathematical knowledge for teaching (Groth, 2007), and it is the dimension that the National Curriculum now emphasises.

Probability–statistical thinking. Conditional probability was the lowest-rated knowledge item (3.45), and the base-rate task produced the lowest correct-response rate (22.3%). This is the clearest instance of partial convergence: both sources identify conditional probabilistic reasoning as an area of need, although the probability block as a whole received a moderate self-rating.

Statistical modelling. The statistics block received one of the highest self-ratings (3.75), and statistical modelling produced the highest objective index (50.2%). This indicates relative consistency between the two sources, but it does not indicate strong competence: the index remained at approximately half of the maximum score, and the two values are measured on different scales in different samples.

Professional-reflexive activity. The self-assessment and objective values in this domain are not numerically comparable. The evidence is instead read for its direction: teaching skill was the lowest-rated survey block, modern methodological approaches and digital tools received lower ratings, the open-ended responses requested methodology, digital tools, and practical training, and the objective pedagogical evidence yielded a derived index of 45.0%. Together these sources indicate partial convergence in the direction of CPD needs, particularly in explaining pupils' reasoning and justifying pedagogical decisions.

3. Comparison with previous research

The low correct-response rates on the base-rate, truncated-axis, and correlation–causation tasks are consistent with international findings on intuitive probabilistic judgement, base-rate neglect, and the interpretation of statistical claims (Batanero et al., 2011; Batanero & Chernoff, 2018; Batanero & Álvarez-Arroyo, 2024), and with Gal's (2002) emphasis on the critical reading of statistical messages. The domain-level divergence between self-assessment and objective evidence is consistent with research showing that self-assessment is an imperfect guide to competence (Dunning et al., 2004), although the present design does not permit individual-level conclusions.

Prior stochastic professional development was associated with higher self-assessed competence. Because the design is cross-sectional and observational, this association may reflect unmeasured factors such as motivation or prior

preparation, and it concerns self-assessment rather than objective performance. No statistically significant association between teaching experience and self-assessment was detected in this sample. One possible interpretation is that, because stochastic content has only recently been emphasised in school mathematics, experience does not automatically substitute for targeted preparation; this is consistent with an andragogical view of experience as a learning resource that must be linked to new content (Knowles, 1980), but it remains an interpretation rather than causal evidence.

4. Implications for continuous professional development

The findings point to content priorities for CPD rather than to a general increase in training time. The priorities derived from the evidence are the interpretation of data, the critical reading of statistical displays, sampling and inference, conditional probability and base-rate reasoning, and the distinction between correlation and causation. The open-ended responses, in which practical, real-life examples (34), more time (19), methodology (9), digital tools and distance formats (9), and textbooks and resources (7) were requested, and the preference for practical training and master classes suggest that these priorities should be addressed through authentic real-life data, digital simulation and visualisation, analysis of pupils' misconceptions and reasoning, and practice-based professional learning. The school-based CPD forms introduced in 2024 offer a suitable vehicle if organised according to the features of effective professional development: content focus, active learning, coherence, collective participation, and sufficient duration (Desimone, 2009).

5. Limitations

Several limitations should be considered. First, the survey and the test were administered to two independent samples, so no individual-level comparison was possible, and the mixed-methods comparison is domain-level rather than individual-level. Second, both samples were convenience samples; the survey sample was drawn from three regions and Tashkent city, background characteristics of the test sample are not reported, and neither sample is nationally representative. Third, the cross-sectional design does not permit causal inference. Fourth, the open-response tasks were scored by a single rater, so inter-rater reliability could not be estimated and the reliability of these scores should

be interpreted with caution; the number of open-response tasks and the rubric dimensions are also not reported. Fifth, the test was not piloted separately. Sixth, the performance bands are descriptive rather than derived from standard setting. Seventh, the objective test does not directly measure classroom practice. Finally, the Component IV index was derived from the equal-weight identity rather than reported directly as an aggregate raw score, and it therefore carries a rounding tolerance of approximately ± 0.3 percentage points.

6. Future research

Future studies should administer both instruments to the same teachers to allow individual-level calibration analysis; use stratified or nationally representative samples; involve multiple raters and report inter-rater reliability for the pedagogical tasks; set performance standards through formal standard-setting procedures; include classroom observation; and adopt longitudinal and quasi-experimental designs, including pre- and post-measures, to evaluate targeted CPD interventions.

CONCLUSION

Teachers reported moderately favourable stochastic competence, whereas objective performance was substantially lower in the assessed domains. Data analysis and interpretation was the weakest quantitatively assessed component, and conditional probability and base-rate reasoning were also problematic. Because the two sources diverged in data interpretation, self-assessment alone may not adequately identify all CPD needs. Combining self-assessment with objective diagnostic evidence provides a stronger basis for CPD planning, and practice-based professional development should target the specific reasoning difficulties identified in this study.

REFERENCES

1. Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, 59(5), 389–407. <https://doi.org/10.1177/0022487108324554>
2. Batanero, C., & Álvarez-Arroyo, R. (2024). Teaching and learning of probability. *ZDM – Mathematics Education*, 56(1), 5–17. <https://doi.org/10.1007/s11858-023-01511-5>
3. Batanero, C., Burrill, G., & Reading, C. (Eds.). (2011). *Teaching statistics in school mathematics: Challenges for teaching and teacher education*. Springer. <https://doi.org/10.1007/978-94-007-1131-0>
4. Batanero, C., & Chernoff, E. J. (Eds.). (2018). *Teaching and learning stochastics: Advances in probability education research*. Springer. <https://doi.org/10.1007/978-3-319-72871-1>
5. Ben-Zvi, D., & Garfield, J. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning and thinking*. Kluwer. <https://doi.org/10.1007/1-4020-2278-6>
6. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
7. Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38(3), 181–199. <https://doi.org/10.3102/0013189X08331140>
8. Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5(3), 69–106. <https://doi.org/10.1111/j.1529-1006.2004.00018.x>
9. Feters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs: Principles and practices. *Health Services Research*, 48(6, Pt. 2), 2134–2156. <https://doi.org/10.1111/1475-6773.12117>
10. Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, 70(1), 1–25. <https://doi.org/10.1111/j.1751-5823.2002.tb00336.x>
11. Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education*, 38(5), 427–437.
12. Knowles, M. S. (1980). *The modern practice of adult education: From pedagogy to andragogy*. Cambridge Adult Education.

13. OECD. (2023). PISA 2022 results (Volume I): The state of learning and equity in education. OECD Publishing. <https://doi.org/10.1787/53f23881-en>
14. President of the Republic of Uzbekistan. (2022). Decree No. PF-134 of 11 May 2022 on approving the national programme for the development of school education in 2022–2026 [in Uzbek]. <https://lex.uz/docs/-6008663>
15. President of the Republic of Uzbekistan. (2024). Resolution No. PQ-231 of 21 June 2024 on additional measures to improve the system of continuous professional development of employees of preschool and school education organisations [in Uzbek]. <https://lex.uz/uz/docs/-6981085>
16. Selyutin, V. D. (2001). Nauchnye osnovy metodicheskoy gotovnosti uchitelya matematiki k obucheniyu shkol'nikov stokhastike [Scientific foundations of mathematics teachers' methodological readiness to teach stochastics to schoolchildren] [Doctoral dissertation]. [in Russian]
17. Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15(2), 4–14. <https://doi.org/10.3102/0013189X015002004>
18. von Davier, M., Kennedy, A., Reynolds, K., Fishbein, B., Khorramdel, L., Aldrich, C., Bookbinder, A., Bezirhan, U., & Yin, L. (2024). TIMSS 2023 international results in mathematics and science. TIMSS & PIRLS International Study Center, Boston College. <https://doi.org/10.6017/lse.tpisc.timss.rs6460>
19. Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. International Statistical Review, 67(3), 223–248. <https://doi.org/10.2307/1403699>
20. Xonqulov, U. X. (2018). Akademik litseylarda matematikaning stoxastika yo'nalishi elementlarini o'qitishning pedagogik imkoniyatlarini takomillashtirish [Improving the pedagogical possibilities of teaching elements of the stochastic strand of mathematics in academic lyceums] [Abstract of PhD dissertation in pedagogical sciences]. Tashkent. [in Uzbek]