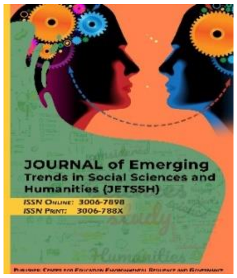

Journal of Emerging Trends in Social Sciences and Humanities

A Double-Blind, Peer-Reviewed, HEC recognized Y-category Research Journal

E-ISSN: 3006-7898 **P-ISSN:** 3006-788X

Volume 3, Issue 3, 2026, 105-116

Perceived Versus Actual Higher Order Thinking Skills: Evidence of a Calibration Gap Among Secondary School Biology Students in Punjab, Pakistan

Rubab Ayyaz[1] Dr. Shumaila Shahzad[2] Dr. Bushra Naoreen[3]

[1] PhD Scholar, Department of Education, Government College University, Faisalabad, Punjab, Pakistan.

[2] Associate Professor, Department of Education, Government College University, Faisalabad, Punjab, Pakistan.

[3] Associate Professor, Department of Education, Government College University, Faisalabad, Punjab, Pakistan.

Corresponding Author: Hina Gull, **Email:** shumailashahzad@gcuf.edu.pk

Article Information

Received:

2026-06-10

Revised:

2026-07-15

Accepted:

2026-08-08

Keywords

Higher Order Thinking Skills; Self-assessment Accuracy; Metacognitive Calibration; Revised Bloom's Taxonomy; Biology Education.

ABSTRACT

Higher order thinking skills (HOTS) occupy the top three levels of the Revised Bloom's Taxonomy and are widely treated as the central aim of contemporary science education. Yet educational practice frequently relies on what students say about their own thinking, on the assumption that self-report tracks capability. This study tested that assumption. Using a cross-sectional survey design, data were collected from 250 Grade 9 Biology students selected through multistage sampling from public-sector secondary schools in four localities of Punjab, Pakistan (Lahore and Multan as urban sites; Depalpur and Bhowana as rural sites). Perceived HOTS was measured with a 20-item, five-point Likert questionnaire covering four dimensions (analysis and application, evaluation, creation, and critical thinking), and actual HOTS was measured with a five-item open-ended achievement test in Biology scored against a four-criterion analytic rubric (score range 20–80). Students rated themselves comparatively highly ($M = 74.70$ of 100, $SD = 8.56$) but performed only moderately on the test ($M = 45.98$ of 80, $SD = 9.18$). A Pearson product-moment correlation found no significant relationship between the two measures, $r(248) = -.043$, $p = .495$, 95% CI [−.17, .08], with shared variance below 1/5 of 1%. The results show that there exists a miscalibration problem in that the students' level of confidence does not provide any useful information about their reasoning. Implications are provided for assessment policy, metacognitive training, and self-reporting in scientific education research.

Publisher: **Institute for Educating Environmental Resilience and Governance**

1. Introduction

Analysing, evaluating, and creating abilities are no longer seen as an adornment in education but rather as its end goal. Ever since Bloom (1956) created the hierarchical arrangement of the cognitive domains and Anderson & Krathwohl (2001) redesigned the hierarchy by placing the cognitive process domains on the vertical axis and the knowledge domains horizontally, the top three domains: Analysing, Evaluating, and Creating have all been classified under HOTS (higher order thinking skills). Brookhart (2010) describes higher order thinking in three related ways: transfer, critical thinking, and problem solving. The common thread linking all three descriptions is that the individual must go beyond the process of retrieval to deal with problems that go beyond learned behavior (Lewis & Smith, 2009).

These skills find natural ground in biology. Life systems are relational, multidimensional, and dynamic; studying life systems thus involves following up structural-functional linkages, considering alternative explanations, analyzing information, and arguing from evidence. It has been acknowledged in this regard by the curriculum frameworks at national level in Pakistan. The National Curriculum for Biology (Ministry of Education, 2006) specifically aims at shifting from assessing knowledge reproduction to problem solving and competency, while the National Education Policy (Ministry of Education, 2009) advocates more scope for independent study, inquiry, critical thinking, and collaboration. Whether the classroom practice has matched the policy is another question. The literature on education in Pakistan has time and again highlighted a “testing culture” that encourages reproduction (Rehmani, 2003; Mahmood, 2017), curricula that emphasize recall (Mahmood, 2010), and teacher training programs that do not prepare teachers to develop analytical thinking skills (Cassum et al., 2013; Dean, 2005).

In this light, there is one measurement issue that does not receive the attention it deserves. In measuring whether students think at high levels, education systems have resorted to asking the learners themselves. Self-report measures are cheap and efficient to use and possess an immediate attractiveness since what else would anyone know better? However, research in the psychology of self-assessment has cautioned for some time now that humans are poor judges of their competence and that the error is not random but systematic (Kruger & Dunning, 1999; Dunning et al., 2004). In the case where the self-rated higher order thinking skills of students fail to correspond with their actual higher-order thinking skills, perception measures will not be able to replace performance data, and inferences based on them will be unreliable.

This paper deals with the above question head-on. Using data collected from 250 Grade 9 Biology students across four localities in Punjab, this study examines the connection between the perceptions of students regarding their higher-order thinking skills and their actual performance on an open-ended, rubric-based achievement test. The study is deliberately narrow in focus: it is not an evaluation of instruction, nor of textbooks, but a test of the correspondence between two ways of measuring the same construct. Its contribution is to supply, for a context in which such evidence is scarce, a quantitative estimate of how much confidence tells us about competence — and to show that in this sample the answer is close to nothing.

1.1 Research Questions and Hypothesis

The study was guided by three questions:

- **RQ1.** What is the perceived level of higher-order thinking skills among secondary school Biology students?
- **RQ2.** What is the actual level of higher order thinking skills among the same students, as measured by a HOTS achievement test?
- **RQ3.** Is there a statistically significant relationship between students' perceived and actual higher order thinking skills?
- The null hypothesis tested was H0: there is no significant relationship between students' perceived and actual levels of higher order thinking skills in Biology at the secondary level.

2. Review of Related Literature

2.1 Higher Order Thinking Skills and the Revised Taxonomy

The Revised Bloom’s Taxonomy (Anderson & Krathwohl, 2001) reformulated the original categories as verbs, reorganized the two uppermost levels to make "creating" precede "evaluating," and combined the dimension of cognitive processes with the dimension of knowledge. Its lasting value comes not from the exact order of the categories but from their practical differentiation, which allows us to define a task based on the level of cognitive challenge it poses to a learner and, hence, audit the curriculum, book, or test for its challenge profile. Analysis, evaluation, and creation have traditionally been categorized as HOTS (Moore & Stanley, 2010), whereas remembering, understanding, and applying have been considered low-order thinking skills (LOTS).

In empirical studies utilizing this approach, it is continually found that instructional materials in many systems are disproportionately distributed at the lower levels. Chiappetta and Fillman (2007) documented the imbalance in secondary biology textbooks; Zohar and Dori (2003) showed that the widespread assumption that lower-achieving students cannot profit from HOTS-oriented instruction is not supported by evidence; and Hugerat and Kortam (2014) demonstrated measurable gains in higher-order thinking when science was taught through inquiry. Within Pakistan, textbook analyses have reported similar imbalances (Mahmood, 2010; Mahmood et al., 2009), and the parent study from which the present paper is drawn found that only about 22% of exercise questions in the Grade IX Punjab Curriculum and Textbook Board Biology textbook addressed the three higher levels.

2.2 Self-Perception, Self-Efficacy, and Metacognitive Calibration

Students' beliefs about their own capability are not epiphenomenal. Self-efficacy beliefs shape task choice, persistence, and effort (Bandura, 1997), and self-regulated learners rely on judgements of their own understanding to decide what to study and when to stop (Zimmerman, 2002). Metacognition — knowledge and monitoring of one's own cognitive processes (Flavell, 1979) — supplies the machinery by which such judgements are made.

The accuracy of those judgements is termed calibration, and the literature on calibration is sobering. Kruger and Dunning (1999) demonstrated that individuals with the weakest skills in a domain tend to overestimate their standing most severely, precisely because the competence needed to perform well overlaps with the competence needed to recognise poor performance. Dunning, Heath & Suls (2004), after reviewing the evidence from the domains of health, education, and work, argue that faulty self-assessment appears to be the rule, not the exception. Specifically, within the context of education, Boud & Falchikov (1989), and Panadero, Brown and Strijbos (2016) found that the relationship between student self-assessment and other measures tends to be weak and highly variable depending on the task, subject domain, and expertise level, while Brown & Harris (2013) warn that self-assessment is justified mainly as an instrument of formative reflection, not measurement.

Two features of that literature matter for the present study. Firstly, calibration is poorer when the task performed is complex and ill-structured compared to well-structured tasks, as the criteria for a good response become less obvious to the novice. Higher-order thinking tasks are precisely of this kind. First, calibration is based on feedback: where the test system provides consistent and clear feedback regarding the level of quality of reasoning, the student's judgement improves; where the test system rewards reproduction, the student gets feedback about his/her memory and not reasoning.

2.3 The Pakistani Context

There are several aspects of the schooling system in Pakistan that make the perception-performance discrepancy a pertinent issue. The exam system continues to be dominated by recall-based tests, which influence not only the curriculum but also how students perceive their performance skills (Rehmani, 2003; Siddiqui, 2017). Textbooks remain the most important learning material, even the sole one at times, while questions within textbooks focus mainly on recall (Mahmood, 2010). Preparation of teachers in critical thinking instruction is uneven (Cassum et al., 2013; Mahmood, 2017), while reform literature continues to outstrip classroom practice (Shah, 2021).

This literature has contained work on the concepts of critical thinking held by teachers, the cognitive level of curriculum content, and the effect of specific instructional approaches. What has not yet been explored, to the best of the authors' knowledge, is whether there is a correspondence between Pakistani secondary students' self-report of higher-order thinking and their actual higher-order thinking skills. This gap is where the present study takes its point of departure. On the practical side, self-report measures are increasingly being used in schools' self-assessment as well as research, where performance tests are expensive to administer. On the theoretical side, a system that offers few opportunities for feedback on reasoning ability makes an ideal laboratory for such a study.

3. Research Methodology

3.1 Research Design

The study employed a quantitative, cross-sectional correlational design. Two instruments were administered to the same students in a single data collection window: a perception questionnaire measuring self-reported HOTS, and an achievement test measuring demonstrated HOTS. Because both variables were measured on continuous scales and the aim was to establish the strength and direction of their association rather than to manipulate any variable, the Pearson product-moment correlation was the primary analytic technique, supplemented by descriptive statistics and independent-samples comparisons on demographic variables. The design is correlational and therefore supports no causal inference in either direction.

3.2 Population and Sample

The target population comprised Grade 9 students studying Biology in public sector secondary schools in Punjab, Pakistan. A multistage sampling procedure was used, a technique appropriate where the population is large and geographically dispersed (Cohen et al., 2018; Creswell, 2014).

At the first stage, four localities were selected purposively to represent contrasting socio-educational settings: Lahore and Multan as urban sites, and Depalpur and Bhowana as rural sites. At the second stage, public sector schools offering Biology at the secondary level under the Punjab Curriculum and Textbook Board syllabus were identified within each locality. At the third stage, students were drawn from school enrolment lists by simple random sampling, so that every eligible student had an equal probability of selection.

The final sample comprised 250 students: 101 males (40.4%) and 149 females (59.6%); 150 urban (60.0%) and 100 rural (40.0%). Because the first stage was purposive, the sample is not a probability sample of Punjab as a whole, and this constrains generalisation (see Section 7).

3.3 Instruments

HOTS Perception Questionnaire. Perceived higher-order thinking was measured with a researcher-developed 20-item questionnaire organised into four five-item dimensions: analysis and application, evaluation, creation, and critical thinking. Each item was rated on a five-point confidence scale anchored at 1 = not confident, 2 = somewhat confident, 3 = moderately confident, 4 = confident, and 5 = very confident. Dimension scores therefore range from 5 to 25 and the total perception score from 20 to 100. Items were worded to refer specifically to Biology learning (for example, analysing biological processes, judging the adequacy of a scientific explanation, or generating an alternative solution to a biological problem) rather than to thinking in general.

HOTS Achievement Test. Actual higher order thinking was measured with a researcher-developed test of five open-ended Biology questions drawn from the Grade 9 syllabus and targeted at the analysing, evaluating, and creating levels of the Revised Taxonomy. The items required students to compare arteries, veins, and capillaries; to explain and justify the characterization of the heart as a double pump; to interpret quantitative observations in biological investigation; to classify organisms across kingdoms and defend the classification; and to explain the significance of interphase in the cell cycle. Open-ended formats were chosen because selected-response items constrain the reasoning that can be displayed (Nitko & Brookhart, 2014).

Responses were scored with an analytic rubric comprising four criteria — understanding and knowledge, analysis and application, evaluation and justification, and creativity and synthesis — each rated from 1 (basic) to 4 (advanced). Each question therefore carried a maximum of 16 points and the test a maximum of 80, with a floor of 20. Scoring the quality of reasoning rather than the correctness of recalled content was central to the construct being measured.

3.4 Validity and Reliability

Content validity for both instruments was established through expert review: practising Biology teachers and educational measurement specialists examined each item for alignment with the intended cognitive level, curricular coverage, and clarity of language, and items were revised in light of their comments. A pilot study with 50 secondary students not included in the main sample was conducted to check item clarity, identify ambiguous wording, and estimate reliability. Internal consistency of the perception questionnaire was examined using Cronbach's alpha against the conventional criterion of ≥ .70 (Nunnally, 1978). For the achievement test, inter-rater and intra-rater agreement on the rubric were examined to establish scoring consistency, since rubric-based scoring of open responses depends on the reproducibility of judgement.

3.5 Data Collection and Ethics

Written permission was obtained from school authorities before data collection. Students were informed of the purpose of the study, told that participation was voluntary and that they could decline without consequence, and assured that responses were confidential and would be used solely for academic research. Both instruments were administered under supervision in a single sitting, with the perception questionnaire completed first so that test performance could not colour self-ratings. No identifying information was retained in the analytic dataset.

3.6 Data Analysis

Data were analysed in IBM SPSS Statistics. Descriptive statistics (means, standard deviations, minima, maxima) summarised both measures. The Pearson product-moment correlation coefficient was computed between total perception and total achievement scores, with significance evaluated at $\alpha = .05$ (two-tailed). Independent samples t-tests were conducted for differences due to gender and urban/rural differences between the two measures, while One-Way Analyses of Variance were done for differences among the four localities. In the case of the present paper, 95% confidence intervals for the correlation coefficient were found using Fisher's r-to-z transformation, while Cohen's d and eta squared values were calculated from the test statistics provided.

4. Results

4.1 Perceived Higher Order Thinking Skills (RQ1)

Descriptive statistics for the perception questionnaire are provided in Table 1. The students ranked themselves quite high on all four dimensions.

Table 1: *Descriptive Statistics for Students' Perceived Higher Order Thinking Skills (N = 250)*

Dimension	Possible range	M	SD	M / item
Analysis and application	5–25	18.45	2.27	3.69
Evaluation	5–25	18.52	2.59	3.70
Creation	5–25	18.58	2.60	3.72
Critical thinking	5–25	19.15	2.56	3.83
Total perception score	**20–100**	74.70	8.56	3.74

Note. Per-item means are computed as the dimension mean divided by the number of items in that dimension and are shown to aid interpretation against the five-point response scale.

A total perception mean score of 74.70 implies an average item score of 3.74 out of five, lying between "moderately confident" and "confident". Higher self-assessments were recorded for critical thinking ($M = 19.15$), while analysis and application received the lowest scores ($M = 18.45$); however, there was little difference in the scores between dimensions – 0.70 scale points, less than one-third of a standard deviation – suggesting that students did not perceive marked differences in their abilities between higher-order thinking skills. A standard deviation of 8.56 from

20-100 implies moderate variation; the group was not equally confident, although low self-scores were rare.

4.2 Actual Higher Order Thinking Skills (RQ2)

Achievement test descriptive statistics are shown in Table 2 below.

Table 2: *Descriptive Statistics for HOTS Achievement Test Scores (N = 250)*

Item (dominant cognitive demand)	Min	Max	M	SD
Q1. Artery–vein–capillary comparison (analyse)	5	16	10.91	2.28
Q2. Heart as a double pump (analyse/evaluate)	4	14	9.21	2.13
Q3. Quantitative observations (evaluate)	4	15	8.66	2.19
Q4. Kingdom classification (analyse/justify)	4	15	8.68	2.36
Q5. Significance of interphase (evaluate/create)	4	15	8.52	2.75
Total HOTS achievement score	21	69	45.98	9.18

Note. Each item was scored on four rubric criteria rated 1–4, giving an item maximum of 16 and a test range of 20–80.

The average of 45.98 marks out of 80 reflects 57.5% of available points, which is equivalent to the average rating of about 2.3 points out of four — in between "developing" and "proficient" categories. Therefore, the performance was moderate but not strong. In the test, the comparative item (Q1, M = 10.91) was performed considerably better than all others; students demonstrated the best results when the question implied an explicit framework for comparing information based on the material that was presented comparatively in the textbook. The lowest results were obtained from the students for Q5 (M = 8.52), in which students had to explain the significance of interphase in the context of the whole cycle, and for Q3 (M = 8.66), in which students had to analyze quantitative observations. These questions imply evaluative reasoning and synthesis, which are not represented in the instructional materials for the students.

The difference between the two measures can be seen from the descriptive statistics. As a ratio of the total scale length, the students' self-reporting was around 74% on the perception scale and achieved only 58% on the achievement test. The two scales are not easily comparable to each other, and this should not be taken to indicate overestimation by a constant margin. However, it sets up the main question of the paper: Are the students who self-reported highly the same students that did well?

4.3 The Relationship Between Perceived and Actual HOTS (RQ3)

A Pearson product-moment correlation was calculated for total perception scores and total achievement scores. Results are presented in Table 3.

Table 3: *Pearson Correlation Between Perceived and Actual Higher Order Thinking Skills*

Variables	N	r	p (2-tailed)	95% CI for r
Perceived HOTS × Actual HOTS	250	−.043	.495	[−.17, .08]

Note. Confidence interval computed by Fisher’s r-to-z transformation. Coefficient of determination r^2 = .0018.

The correlation coefficient was negative, small in size, and not statistically significant, r (248) = -.043, p = .495. The null hypothesis could not be rejected. There are three aspects to this finding which require highlighting.

First, the coefficient of determination is .0018, meaning that perception explains less than one-fifth of one percent of the variance in academic performance. This is not a small effect; it is essentially the absence of an effect.

Secondly, the output does not simply reflect an inability to achieve significance due to lack of statistical power. At N = 250, the test possessed about 80% power to find a correlation as low as r = .18 at the α = .05 level. The 95% confidence interval, [−.17, .08], excludes every value that would be of practical use for prediction; even the most favourable endpoint of the interval corresponds to less than 1% shared variance. The finding is therefore an informative null rather than an inconclusive one.

Third, the negative sign should not be over-interpreted as evidence that confidence impairs performance. Its origin is examined in Section 4.5.

4.4 Group Comparisons on the Two Measures

The demographic comparisons reported in the parent study bear directly on the interpretation of the correlation and are summarised in Table 4.

Table 4: *Summary of Group Comparisons on Perceived and Actual HOTS*

Comparison	Measure	Group means	Test	p	Effect size
Gender	Perception	M 72.35 / F 76.29	t(248) = −3.66	< .001	d = −.47
Gender	Achievement	M 48.51 / F 44.26	t(248) = 3.69	< .001	d = .48
Residence	Perception	U 75.29 / R 73.81	t(248) = 1.34	.182	d = .17
Residence	Achievement	U 46.09 / R 45.81	t(248) = 0.24	.812	d = .03
City	Perception	see note	F(3, 246) = 6.54	< .001	η^2 = .074
City	Achievement	see note	F(3, 246) = 4.46	.005	η^2 = .052

Note. M = male, F = female, U = urban, R = rural. City perception means: Lahore 78.28, Bhowana 76.14, Multan 73.79, Depalpur 71.48. Turkey post-hoc comparisons indicated Lahore > Multan and Bhowana > Depalpur on perception, and Lahore > Multan and Lahore > Depalpur on achievement. Effect sizes were computed by the present authors from the reported test statistics.

The pattern is instructive. On the two variables where a difference was detected, gender and locality, the two measures do not merely fail to agree in magnitude — on gender they point in opposite directions. Female students rated themselves significantly higher than male students, with a moderate effect size, while scoring significantly lower on the achievement test, with an almost identical effect size in the opposite direction. Residence, by contrast, produced no significant difference on either measure, and locality produced significant differences on both, though of modest magnitude (approximately 7% and 5% of variance respectively).

4.5 Decomposing the Negative Correlation

Because the gender differences run in opposite directions on the two measures, the pooled sample contains a compositional effect: the group that scores higher on perception is also the group that scores lower on achievement, which imposes a negative component on the overall covariance irrespective of any individual-level association.

This component can be quantified. The total covariance implied by the reported statistics is r × SD_perception × SD_achievement = −.043 × 8.56 × 9.18 = −3.38. The between-gender component, computed from the group means and group sizes, is −4.03 — larger in magnitude than the total. The residual within-gender covariance is therefore approximately +0.66, corresponding to a within-gender correlation of roughly +.01.

Put differently, the whole negative aspect of the correlation observed is due to the composition of the sample, and not because of the likelihood of overconfident people doing poorly. Controlling for gender, the relationship between subjective and objective HOTS scores is not negative; it is close to zero. Far from weakening the conclusion reached, this result makes the conclusion stronger since it shows that self-ratings contain no useful information about performance, either overall or within gender categories. The relationship appears low because of the aggregation problem as described by Simpson’s paradox. This particular decomposition is based on published means and is therefore approximate.

5. Discussion

5.1 A Calibration Gap, not a Confidence Deficit

The key finding is that students' evaluations of their HOTS skills are useless when it comes to predicting their HOTS demonstrated abilities. What is important is the fact that the problem does not lie in students' confidence level. They considered themselves to be either confident or very confident in all four aspects of HOTS, even in creation, which is the highest level in the HOTS taxonomy.

This conforms to calibration research findings. Kruger & Dunning (1999) and Dunning et al. (2004) posit that correct self-assessment entails the competence being assessed because the criteria that define a good from a poor analysis are integral to the competence itself. A student who lacks an understanding of the nature of a well-substantiated justification cannot realize that his/her own justification is shallow. According to Panadero et al. (2016) and Boud and Falchikov (1989), self-assessment correlations with external criteria are usually small even in favorable situations; indeed, the current correlation falls at the low end of that distribution, considering the nature of the criterion task.

Another contextually sensitive explanation is the nature of the feedback context. Students judge their level of competence based on the feedback that is available to them; in an assessment system where recall questions dominate the exams, the feedback relates to the students' ability to remember rather than reason. Where a student consistently reproduces textbook material, they consistently receive feedback that tells them that they are succeeding academically and may generalize this feedback to include their reasoning and evaluation abilities – abilities that are seldom tested and thus seldom disconfirmed. According to this interpretation, the problem of calibration is not one of inadequate students but a result of an assessment system that provides no feedback on the abilities in question.

5.2 The Gender Pattern

The most significant result in the group comparisons is that of the opposite gender results, where there was high perception in females and high achievement in males, both having moderate effect size. The results should be interpreted with care rather than confidence.

Some reasons are quite possible. The response style could be systematically different between the groups, whereby one group would be more likely to report high ratings of confidence independent of their actual abilities. Taking an open-response test over a long period of time might be affected by the person's willingness to answer a tough question within the allotted time, rather than being an indicator of the reasoning ability that the test should measure.

The information that is not supported by the data is the notion that either group has a superior ability to engage in higher-order thinking. One achievement test taken at one point in time in one subject cannot justify this statement, and according to general literature, both groups demonstrate the same ability to engage in higher-order thinking under the same conditions. The proper interpretation of the data should be more limited in scope and use: self-evaluation bias does not exist equally among all groups, and therefore, aggregate self-reporting data might misrepresent differences among groups. According to perception data from the school, females would be better thinkers than males.

5.3 Consistency with the Instructional Environment

The average achievement score and the pattern within the test are consistent with prior knowledge of the instructional material. In the parent study, the content analysis revealed that around 78 percent of the Grade IX Biology textbook exercises were related to the lower three categories of the taxonomy, while analyzing constituted 12 percent, evaluating six percent, and creating four percent. Learners did well in the question most similar to the comparison found in the textbook but poorly in the questions demanding evaluative justification and synthesis, which were the areas least covered by the learners' instructional material.

The lack of distinction between urban and rural settings follows similar reasoning. Where the curriculum, textbook, and examination system are standardized within the province, the factors exerting pressure for higher cognitive engagement are uniform, and the effect of setting is minimal. The fact that such distinctions were found among localities but not the urban and rural divide indicates that the variations exist at the level of school/classroom practice, such as pedagogy, questioning strategy, and teacher preparation. This is more practical advice than its alternative, which is out of reach for action due to the vagueness of the setting.

5.4 Implications for the Use of Self-Report Instruments

The clear implication of all of the above is the methodological one: perception questionnaires have become an integral part of educational research and school self-assessments due to their assumed correlation with performance. In this sample, they did not correlate with performance at all. Where the object of interest is capability, self-report cannot stand in for demonstration.

This does not render perception data worthless. Self-efficacy beliefs predict engagement, persistence, and subject choice (Bandura, 1997; Zimmerman, 2002), and are worth measuring in their own right. The error lies in treating them as measures of a different construct. Perception and performance data are complementary: measured together, the discrepancy between them is itself diagnostically valuable, identifying students who are overconfident and at risk of under-preparing, as well as students who are capable but under confident and may avoid demanding work.

6. Implications and Recommendations

Four recommendations follow from the findings.

Make reasoning quality visible to students. Calibration improves when learners receive explicit, criterion-referenced feedback on the quality of their thinking. Sharing analytic rubrics with students before they attempt a task, requiring them to predict their own rubric rating and then compare that prediction against the assigned rating, and discussing exemplar responses at different levels of quality all supply the information that self-judgement currently lacks. This is a low-cost intervention that requires no curricular change.

Rebalance the cognitive demand of textbook exercises and classroom assessment. If under a quarter of textbook questions require analysis, evaluation, or creation, students will have limited opportunity to practise those skills and limited evidence on which to judge themselves. Revision of exercise sets to include comparative analysis, justification of conclusions, interpretation of data, and open problem-solving would address both the achievement and the calibration findings simultaneously.

Embed metacognitive instruction in science teaching. Prompting students to articulate why they chose an approach, what evidence supports their conclusion, and where their reasoning is weakest builds the monitoring capacity that accurate self-assessment presupposes (Flavell, 1979; Zimmerman, 2002). These prompts are part and parcel of inquiry-based learning and problem-based learning rather than something that is additional to such approaches.

Pair subjective and objective measurement in evaluation and research. Thinking skills assessments using institutional data must not rely solely on student self-reports. When both are obtained, the gap between perception and performance must be reported as a separate finding and not averaged out.

7. Limitations

There are a number of limitations associated with these findings. The study uses a cross-sectional and correlational design; hence, it does not make any claims regarding the direction of the influence between perception and performance. The first level of sampling used is purposive, while at the last level the students were randomly chosen, but the four localities are not probability samples of Punjab. Actual HOTS was measured with a five-item test administered on a single occasion in a single subject, which limits content coverage and leaves the estimate vulnerable to occasion-specific factors such as fatigue, unfamiliarity with open-response formats, and time pressure; a longer instrument, or several administered across topics, would yield a more stable estimate. Both instruments were researcher-developed, and while content validity was established through expert review and reliability was examined in a pilot, the specific reliability coefficients should be reported alongside these results for full transparency. The perception and achievement instruments are not measured on a common metric, so the study establishes the absence of association rather than the magnitude of overestimation. Finally, the decomposition in Section 4.5 is computed from published group statistics rather than raw data and should be verified on the original dataset, ideally by computing the correlation separately within each gender group.

8. Conclusion

This study asked whether secondary school Biology students' beliefs about their own higher-order thinking correspond to their demonstrated higher-order thinking. Among 250 Grade 9 students in Punjab, they did not. Students rated themselves between moderately confident and confident across analysis, evaluation, creation, and critical thinking, yet achieved only moderate scores on a rubric-scored open-response test, and the correlation between the two measures was negligible and non-significant, $r(248) = -.043$, $p = .495$. The confidence interval was narrow enough to exclude any association of practical value, making this an informative null result rather than a failure of statistical power. Analysis of the group comparisons further showed that the small negative sign arises from the opposing gender pattern in the two measures rather than from any individual-level relationship.

The practical message is straightforward. Students cannot presently tell how well they think, and in an assessment environment that rarely tests reasoning, there is little reason to expect that they could. Asking them is therefore not a shortcut to knowing. Higher-order thinking must be taught deliberately, assessed directly, and fed back explicitly — and until the feedback loop that would make self-judgement informative is built, perception data should be read as evidence about confidence, and about nothing else.

Conflict of Interest

The authors showed no conflict of interest.

Funding

The authors did not mention any funding for this research.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives: Complete Edition*. Addison Wesley Longman, Inc.

Bandura, A. (1997). *Self-efficacy: The Exercise of Control*. Macmillan.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Handbook I: cognitive domain. *New York: David McKay*, 483-498.

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education*, *18*(5), 529-549.

Brookhart, S. M. (2010). *How to Assess Higher-Order Thinking Skills in Your Classroom*. Ascd.

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *SAGE Handbook of Research on Classroom Assessment* (pp. 367–393). SAGE Publications.

Cassum, S. H., McGrath, J. P., Gul, R. B., Dilshad, A., & Syeda, K. (2013). Multidimensionality of critical thinking: A holistic perspective from multidisciplinary educators in Karachi, Pakistan. *Journal of Nursing Education and Practice*, *3*(7), 9.

Chiappetta, E. L., & Fillman, D. A. (2007). Analysis of five high school biology textbooks used in the United States for inclusion of the nature of science. *International Journal of Science Education*, *29*(15), 1847-1868.

Cohen, L., Manion, L., & Morrison, K. (2009). *Research methods in education* (Vol. 6). London: routledge.

Creswell, J. W., & Creswell, J. D. (2017). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Sage publications.

Dean, B. L. (2005). Citizenship education in Pakistani schools: Problems and possibilities. *International Journal of Citizenship and Teacher Education*, *1*(2), 35.

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.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (1993). *How to Design and Evaluate Research in Education* (Vol. 7). New York: McGraw-hill.

Hugerat, M., & Kortam, N. (2014). Improving higher order thinking skills among freshmen by teaching science through inquiry. *Eurasia Journal of Mathematics, Science and Technology Education*, *10*(5), 447-454.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121.

Lewis, A., & Smith, D. (1993). Defining higher order thinking. *Theory into Practice*, *32*(3), 131-137.

Mahmood, K., Iqbal, M. Z., & Saeed, M. (2009). Textbook evaluation through quality indicators: The case of Pakistan. *Bulletin of Education and Research*, *31*(2), 1-27.

Mahmood, K. (2010). Textbook evaluation in Pakistan: Issue of conformity to the national curriculum guidelines. *Bulletin of Education and Research*, *32*(1).

Mahmood, S. (2017). *Testing the Effectiveness of a Critical Thinking Skills Intervention for Initial Teacher Education Students in Pakistan* (Doctoral dissertation, University of Southampton).

Ministry of Education. (2006). *National Curriculum for Biology: Grades IX–XII*. Government of Pakistan.

Ministry of Education. (2009). *National Education Policy 2009*. Government of Pakistan. https://www.mofept.gov.pk/SiteImage/Policy/NEP_2009.pdf

Moore, B., & Stanley, T. (2010). *Critical thinking and formative assessments: Increasing the Rigor in Your Classroom*. Eye on Education.

Brookhart, S. M., & Nitko, A. J. (2019). *Educational Assessment of Students*. Upper Saddle River, NJ: Pearson.

Nunnally, J. C. (1978). Psychometric Theory 2nd ed: Mcgraw hill book company.

Panadero, E., Brown, G. T., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review*, *28*(4), 803-830.

Pintrich, P. R., & Schunk, D. H. (1996). Motivation in education: Theory, research, and applications. *(No Title)*.

Rehmani, A. (2003). Impact of public examination system on teaching and learning in Pakistan. *International Biannual Newsletter ANTRIEP*, *8*(2), 3-7.

Shah, D. (2021). Examining the disconnect between national curriculum policy and classroom practice in Pakistan. *Education Research International, 2021*, 1–12.

Siddiqui, S. (2016). Education policies in Pakistan: Politics, projections, and practices. *(No Title)*.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, *41*(2), 64-70.

Zohar, A., & Dori, Y. J. (2003). Higher order thinking skills and low-achieving students: Are they mutually exclusive? *The journal of the Learning Sciences*, *12*(2), 145-181.