education sciences

MDPI

Article

Low Inter-Rater Reliability of a High Stakes Performance Assessment of Teacher Candidates

Scott A. Lyness *, Kent Peterson and Kenneth Yates

Rossier School of Education, University of Southern California, Los Angeles, CA 90089, USA; kentpeterson2@gmail.com (K.P.); kennetay@usc.edu (K.Y.)
* Correspondence: lyness@usc.edu

Abstract: The Performance Assessment for California Teachers (PACT) is a high stakes summative assessment that was designed to measure pre-service teacher readiness. We examined the inter-rater reliability (IRR) of trained PACT evaluators who rated 19 candidates. As measured by Cohen's weighted kappa, the overall IRR estimate was 0.17 (poor strength of agreement). IRR estimates ranged from −0.29 (worse than expected by chance) to 0.54 (moderate strength of agreement); all were below the standard of 0.70 for consensus agreement. Follow-up interviews of 10 evaluators revealed possible reasons we observed low IRR, such as departures from established PACT scoring protocol, and lack of, or inconsistent, use of a scoring aid document. Evaluators reported difficulties scoring the materials that candidates submitted, particularly the use of Academic Language. Cognitive Task Analysis (CTA) is suggested as a method to improve IRR in the PACT and other teacher performance assessments such as the edTPA.

Keywords: inter-rater reliability; preservice teacher performance assessment; PACT; edTPA; weighted kappa; cognitive task analysis; qualitative; quantitative

Citation: Lyness, S.A.; Peterson, K.; Yates, K. Low Inter-Rater Reliability of a High Stakes Performance Assessment of Teacher Candidates. *Educ. Sci.* **2021**, *11*, 648. https://doi.org/10.3390/educsci11100648

Academic Editors: Eila Jeronen and James Albright

Received: 25 August 2021
Accepted: 6 October 2021
Published: 18 October 2021

1. Introduction

Successful education systems are complex and require classroom teachers who demonstrate a broad range of competencies [1]. One way to increase the potential of placing competent teachers in classrooms is through the performance assessment of teacher candidates. Novice teacher competency is assessed within teacher preparation programs, and frequently, in the United States, by state-mandated assessments of teacher performance. To effectively assess candidates' readiness to teach, these assessments need to be used in a manner that optimizes transparency, authenticity, validity, and reliability.

Teaching performance assessments (TPAs), such as the Performance Assessment for California Teachers (PACT), were mandated by the California state legislature in 1998 in response to a perceived need for teachers, and teacher education programs, to be held more accountable for the teaching skills and expertise brought to the K-12 classroom. Like exit exams in law and medical school, an exit exam in schools of education was suggested so as to raise the stature of the teaching profession among other professions [2], and to ultimately improve the quality of teacher education programs.

The PACT is a standards-based, performance test of subject-specific content areas. It was reasoned that an authentic, clinical demonstration of an actual teaching segment, as opposed to completing a multiple-choice test, would lead to better teaching, improved teacher education programs, and ultimately, improved learning outcomes in the students that the teachers taught. These are all laudable goals, but more evidence-based research on these outcomes is still pending.

Teacher candidates in California must pass an approved teacher performance assessment (TPA) to be recommended for a preliminary teaching credential. Even though the PACT has been discontinued as a teacher performance assessment at our institution, it has evolved into the edTPA [2,3], which is used nationally at several institutions, including our

own. Both the PACT and edTPA share similar "content, design, and authorship" [4] (p. 73) and originated at the Stanford Center for Assessment, Learning and Equity (SCALE). A major difference between the two TPAs is that the PACT is locally scored, whereas the edTPA is outsourced to evaluators.

Issues and concerns that have been raised about TPAs in the United States [3–6] (e.g., costs, narrowing the teacher education curricula to accommodate the TPA, the considerable stress that TPAs can cause teacher candidates, outsourced scoring, issues of authenticity, and concerns about reliability and validity) have also been raised about TPAs internationally [7–9].

We wanted to examine the inter-rater reliability (IRR) of our candidates' edTPA data, but when we requested double-scored data from the edTPA, we were not given access because the data are proprietary to Pearson Inc. We therefore examined the IRR in our PACT data that was collected in the last year we had data for. We believed this was important for several reasons:

- IRR in teacher performance assessments has not often been reported.
- When IRR has been reported, it has been reported many times as percent agreement, which is considered to be a "common mistake" [10] (p. 4).
- We had the double-scored data; these data are hard to come by now that they are proprietary to corporations such as Pearson.
- Because of the high stakes nature of the TPA (receiving or not receiving your credential), we felt it was important to estimate the IRR in our sample.

Reliability is the degree to which assessments produce stable and consistent results. An assessment cannot be valid if it is not reliable [11]. Reliability is challenging to establish in performance assessments [12] (pp. 222–223) because every teacher candidate portfolio submission for a TPA is unique. As opposed to multiple-choice tests with selected response items, TPAs consist of constructed response items that need to be subjectively scored by evaluators, which introduces a potential source of error. Constructed response items have some of the lowest reliabilities [3] (p. 74). Because of the high stakes nature of teacher performance assessments such as the PACT, reliability is important to establish [13], and is an ongoing process [14]. In the present study, our data provide information about one form of reliability, namely, inter-rater reliability (IRR), which is the degree to which different observers agree on what they are rating [3]. Gitomer et al. [5] report that the main source of error in edTPA scores is typically the rater. For different ways in which IRR can be calculated, and a discussion of their advantages and disadvantages, please see [10,15].

Our review of the literature revealed relatively few studies examining IRR of TPAs. Riggs et al. [16] examined the IRR for the California Teaching Performance Assessment (CalTPA) in a pilot project. The CalTPA consists of four tasks (Subject-Specific Pedagogy, Designing Instruction, Assessing Learning, and the Culminating Teaching Experience), which are used to measure 12 Teacher Performance Expectations. Each task is scored ordinally with either a 1 (low score), 2, 3, or 4 (high score). IRR was computed using intra-class correlations. Of the 20 scores reported, 17 were below 0.40 (poor) and 3 were in the 0.40 to 0.59 (fair) range. The authors concluded that IRR for the CalTPA was "inadequate" [16] (p. 24).

Regarding prior reports of the IRR in the PACT, most have reported percentage agreement (e.g., [17]), even though this has been rejected as an adequate measure [10,18]. Percentage agreement does not take into account chance agreement. Hallgren [10] states that reporting percentage agreement is a common mistake when reporting IRR. To determine the IRR for the PACT, Pecheone and Chung [17] reported data for the first 2 years of the pilot project. For the 2002–2003 pilot study, 163 out of 395 teaching events were double scored for IRR. Inter-rater agreement percentage was 90% (score pairs were exact plus adjacent agreement). For the 2003–2004 pilot study, 203 out of 628 teaching events were double scored for IRR; inter-rater agreement percentage was 91% for an exact plus adjacent agreement.

In a study by Porter [19] and Porter and Jelinek [20], IRR of the PACT was assessed by Cohen's kappa [18] as well as percentage agreement. Kappa is a measure of agreement with

chance agreement accounted for. Overall, they reported a polytomous kappa of 0.33 [19], which corresponds to a "fair" strength of agreement. In this current study, we also examine our PACT data with Cohen's kappa. Even though percentage agreement is an inferior measure of IRR, we also report it so that we can compare our results to prior studies that reported percentage agreement.

Interestingly, although the literature contains a few studies examining the quantitative results of PACTs [17,19,20], a review of the literature did not reveal qualitative studies that investigated how evaluators assess PACTs. This presents a significant gap when attempting to provide context for an IRR estimate. In the present study, we attempted to bridge that gap by first calculating the IRR of PACTs completed by evaluators in one teacher preparation program and then engaging 10 PACT evaluators in follow-up interviews structured to gain a better understanding of the cognitive processes, analyses, and decisions evaluators use when assessing PACTs. Combining current IRR data with findings from in-depth interviews may illuminate the mechanisms that may limit the reliability, usefulness, and fairness of the PACT. Moreover, these findings may identify areas for improving evaluators' ratings and IRR of the PACT as an authentic, valid, and reliable assessment of teacher candidates' qualifications to enter the classroom. In sum, we wanted to answer two research questions:

- What is the inter-rater reliability (IRR) of a sample of PACTs from our teacher preparation program?
- Can evaluator descriptions of how they assessed the sample PACTs illuminate and explain the IRR results?

2. Methods

The PACT assessed pre-service candidates' abilities on five tasks to plan, instruct, assess, reflect, and encourage student use of academic language [17]. Each of these five tasks included two or three rubrics (see Table 1), which were each scored 1 (Fail), 2 (Basic or Pass), 3 (Proficient), or 4 (Advanced). An example of PACT elementary math rubrics can be found in Porter [19] (pp. 124–132). These scores were derived from the evaluation of two types of evidence in PACT submissions: artifacts (e.g., lesson plans, videos, student work samples) were evidence that candidates submitted to show teacher competence, and commentaries (written responses to standardized questions that provided context and rationale for the artifacts submitted).

Table 1. Performance Assessment for California Teachers (PACT) rubrics.

Task	Rubric
Planning	1. Establishing a balanced instructional focus 2. Making content accessible 3. Designing assessments
Instruction	4. Engaging students in learning 5. Monitoring student learning during instruction
Assessment	6. Analyzing student work from an assessment 7. Using assessment to inform teaching 8. Using feedback to promote student learning
Reflection	9. Monitoring student progress 10. Reflecting on learning
Academic Language [a]	11. Understanding language demands and resources 12. Developing students' academic language repertoire

[a] Assessed throughout the teacher performance event.

2.1. PACT Training, Calibration, and Scoring

In the 2014 academic year (the last time we collected PACT data), 22 evaluators assessed 128 PACTs in our teacher preparation program. The evaluators consisted of 17 part-time faculty, four K-12 administrators, and one K-12 classroom teacher. All evaluators participated in a one time, 2-day training that included both conceptual knowledge of the terms and rubrics as well as demonstrations, practice, and feedback on the strategies and

sequence of evaluating candidates' submissions. After their initial training and calibration, PACT evaluators were required to attend annual recalibration events.

Each content area had their own benchmark PACT and the benchmarks changed on a yearly basis. To demonstrate evaluation competence, 6 out of 12 rubrics had to have perfect agreement, along with no instances where there was a disagreement of more than one rubric level. If the calibration standard was not met, the potential PACT evaluator would meet with the PACT coordinator to discuss the scores and review the "Thinking Behind the Rubrics" document [21] until a consensus was reached. This document was available to all evaluators and served as a tool to better understand each rubric through nuanced descriptions of what each rubric meant, the intentions behind the rubrics, and suggestions for what to look for when determining scores.

According to the PACT training protocols, evaluators were instructed to:

- Complete evaluations in order, starting with a review of the background information provided by the candidate describing the student population including geographic location, cultural context, race and ethnicity, and other statistical data such as free and reduced lunch percentages.
- Review the PACT sections in order: Planning, Instruction, Assessment, and Reflection, while addressing use of Academic Language throughout.
- Identify evidence that met the criteria of a rating of 2 first (basic novice teacher competency), and then determine if the evidence was above or below this mark. Evaluators were then trained on evidence that defined a 1 or a 3 in all tasks. There was no specific training for what defines a 4 rating.
- Score the evidence submitted by the candidate without inferring what the candidate might have been thinking or intending.
- Assess PACTs using the rubrics only.
- Take notes as they assess.
- Consistently refer to the "Thinking Behind the Rubrics" document to provide a more in-depth explanation of rubric scores.
- Recognize their own biases.

The results of the PACT assessment were examined to determine the pass/fail rate to comply with state guidelines for additional review. Of the 128 PACTs, 120 passed and eight failed. The eight failing PACTs were double scored to comply with state policy. Double scoring procedures for the eight failed PACTs included an initial reading by the original evaluator, followed by a second reading by another PACT evaluator in the same content area. If these two evaluators agreed that a PACT failed, it was returned to the candidate for remediation. Remediation efforts involved review and generalizable feedback provided by the PACT coordinator and, on occasion, other faculty members. If the first two evaluators disagreed, then a third evaluator was used as a tiebreaker. Typically, this third evaluator was a lead faculty member.

From the remaining 120 passing PACTs, 11 were randomly selected to comply with the state policy of double scoring 15% of the total completed PACTs. Passing candidates were notified and no further action was taken on their PACT submission. Of the 19 double-scored PACTs (eight from the failing category and 11 randomly selected from the passing category), four TPAs were in the multiple subject area and focused on math; fifteen were in the single subject area: history (n = 4), math (2), science (5), and English language arts (4).

We conducted the study in two stages. First, we calculated the IRR for the double-scored PACTs. Second, we interviewed evaluators to examine the methods they used to score the PACTs.

2.2. IRR Calculation

Consensus agreement occurs when reasonable observers come to an agreement on how the criteria of a rubric have been applied to observed phenomena [15]. Stemler [15] notes "a typical guideline found in the literature for evaluating the quality of inter-rater reliability based upon consensus estimates is that they should be 70% or greater." We

examined consensus agreement in two ways. First, we computed Cohen's kappa coefficient of agreement, k [18], which is directly interpretable as the proportion of agreement with chance agreement excluded. Kappa ranges from −1 (perfect disagreement) to 1 (perfect agreement). Table 2 shows the range of kappa and corresponding strength of agreement when we made our original computations [22]. Because the scores for each rubric are ordinal (1 to 4), a weighted Cohen's kappa, k_w, was computed [22,23]. The computation of weighted kappa considers close matches, or how far apart the raters are. For example, if one evaluator assigns a rubric a 2 and another evaluator assigns the same rubric a 3, this is a closer match than one evaluator assigning a 2 and another evaluator assigning a 4. We used linear weighting. Standard errors and 95% CIs for k_w were computed mostly from www.vassarstats.net/kappa.html (accessed on 12 August 2021) [24].

Table 2. Cohen's kappa (k) and corresponding strength of agreement.

k					
−1 to <0	0 to 0.20	>0.20 to 0.40	>0.40 to 0.60	>0.60 to 0.80	>0.80 to 1
Worse than expected by chance	Poor	Fair	Moderate	Good	Very good to perfect agreement

Second, even though it is considered to be a mistake to report IRR with percentage agreement [10], we wanted to compare our results using this method to the way IRR of the PACT has typically been reported. Therefore, we computed percentage agreement as exact agreement (when both evaluators assign the same score), or as exact plus adjacent agreement (when the paired evaluators assign the same score or are within 1 point of the score). As stated earlier, the problem with using percentage agreement is that it does not take into account agreement by chance and, therefore, inflates the level of agreement.

2.3. Interviews

Following the University's Institutional Review Board approval, we solicited all 22 PACT evaluators to participate in semi-structured interviews. From the 22, ten agreed to participate: six were part-time faculty, three were school administrators, and one was a classroom teacher. The interviews (see Appendix A) were conducted by either telephone or in the interviewer's online classroom, lasted approximately 30 min, and were transcribed. The transcripts were independently analyzed by two researchers (S.A.L. and K.P.). The interview text was uploaded into Atlas.ti, a qualitative data analysis program [25]. A document was made for each interview question and each evaluator's response was determined. Responses for each interview question were partitioned into categories with similar responses, and frequency counts were made. A consensus was reached, and the patterns of the data that emerged from the frequency analysis are reported.

3. Results

We first present IRR, computed by weighted kappa and percentage agreement, followed by the interview results from 10 of the evaluators. As previously stated, 19 candidates were independently rated by two evaluators on the 12 rubrics in Table 1. Each rubric was rated from 1 (Fail) to 4 (Advanced). Table 3 shows the raw double scores for each of the 19 candidates, so that the reader may replicate the results reported here. Note that 4s were rarely given (8/456, <2%).

A 4 × 4 matrix was created for each candidate and the paired rating for both evaluators was put in the appropriate cell. As an example, the first candidate's matrix is shown in Table 4. Cohen's weighted kappa and percentage agreement were computed from this matrix for each candidate.

Table 3. Raw double-scored data used to compute inter-rater reliability.

Candidate	Evaluator	P1	P2	P3	I4	I5	A6	A7	A8	R9	R10	AL11	AL12
1	1	3	2	2	2	1	1	1	2	1	1	1	2
	2	3	3	2	2	1	2	1	2	2	1	2	2
2	1	3	3	3	2	2	2	3	2	3	2	2	3
	2	3	4	3	3	4	3	3	4	3	3	3	3
3	1	4	3	4	3	2	3	3	4	3	2	3	2
	2	3	3	3	3	2	4	2	2	3	3	2	2
4	1	2	2	2	3	2	3	3	3	3	3	3	2
	2	3	2	2	3	2	2	2	3	2	2	2	2
5	1	3	2	2	2	2	2	2	2	2	2	2	2
	2	3	3	2	2	2	1	1	1	3	3	2	2
6	1	2	1	1	2	1	1	1	1	2	1	1	1
	2	3	2	2	2	1	2	1	1	2	2	1	1
7	1	2	3	2	3	2	2	2	2	2	2	2	2
	2	2	2	2	1	1	2	1	2	2	2	2	2
8	1	2	2	2	2	2	2	2	2	2	2	1	1
	2	2	2	1	1	1	2	1	2	2	2	2	2
9	1	2	2	2	2	1	2	1	1	2	2	2	2
	2	2	2	2	1	1	2	2	2	2	2	2	2
10	1	2	2	3	1	1	2	1	1	2	2	2	2
	2	2	2	2	1	1	2	2	2	2	2	2	1
11	1	2	2	3	3	2	3	2	3	3	2	2	2
	2	3	2	3	2	1	2	2	3	2	2	2	2
12	1	3	3	2	3	2	3	2	2	2	3	3	2
	2	3	3	2	3	3	3	3	2	3	3	2	2
13	1	2	3	2	2	3	2	2	1	1	2	2	2
	2	3	3	2	2	2	2	2	2	3	2	2	2
14	1	2	2	1	2	2	1	2	1	2	2	2	1
	2	1	1	1	1	1	1	1	1	1	2	1	1
15	1	3	3	3	3	2	3	3	4	2	3	3	2
	2	3	3	3	2	2	2	3	2	2	2	2	3
16	1	1	2	2	1	1	1	1	2	1	1	2	2
	2	3	2	3	1	1	1	1	1	1	2	1	1
17	1	2	3	3	2	2	3	2	1	1	2	2	2
	2	2	2	2	2	2	2	2	1	1	2	1	1
18	1	3	3	3	3	2	2	3	3	3	2	2	3
	2	2	2	2	2	2	2	2	2	2	2	2	2
19	1	2	3	3	1	1	2	2	1	2	2	2	2
	2	3	3	2	2	1	2	2	2	2	2	2	2

Note. P = Planning rubrics, I = Instruction rubrics, A = Assessment rubrics, R = Reflection rubrics, AL = Academic Language rubrics.

Table 4. Cross tabulation showing paired ratings for Candidate 1.

	Evaluator 2				
	PACT Score				
Evaluator 1	**1**	**2**	**3**	**4**	**Total**
1	3	3	0	0	6
2	0	4	1	0	5
3	0	0	1	0	1
4	0	0	0	0	0
Total	3	7	2	0	12

Note. K_w = 0.54. The diagonal shows exact agreement = 8/12 = 66.7%. Off-diagonals show disagreement. Exact + adjacent agreement results in 100% agreement.

The IRR results are reported in Table 5. Eight candidates initially failed the PACT. Of these eight, one candidate (Candidate 19) was rated to pass by a second evaluator, and subsequently in a tiebreaker, by a third evaluator. Note that three candidates (Candidates 5, 7, and 17) initially passed the PACT; these were randomly chosen to be re-scored and

were rated as failing the PACT by a second evaluator. Candidate 17 was mistakenly rated by the same evaluator twice where the candidate first passed, and then failed the PACT; note that the weighted kappa (0.35) for Candidate 17 is only fair and percentage agreement (exact) between the two ratings is less than acceptable (58.3%).

Table 5. Candidates' inter-rater reliabilities: Cohen's weighted kappa (with standard error, 95% confidence interval, strength of agreement), and percentage agreement.

Candidate—Initial, Final Pass (P) or Fail (F)	Cohen's Weighted Kappa, k_w	Standard Error	95% CI	Kappa Strength of Agreement	Percentage Agreement (Exact)	Percentage Agreement (Exact + Adjacent)
1—F, F	0.54	0.19	0.17 to 0.90	Moderate	66.7	100
2—P, P	0	—	—	Poor	41.7	83.3
3—P, P	0.11	0.18	−0.24 to 0.46	Poor	41.7	91.7
4—P, P	0.08	0.22	−0.36 to 0.51	Poor	50	100
5—P, F	0.18	0.15	−0.12 to 0.49	Poor	50	100
6—F, F	0.33	0.16	0.02 to 0.65	Fair	58.3	100
7—P, F	0	—	—	Poor	66.7	91.7
8—F, F	−0.29	—	—	Worse than expected by chance	50	100
9—F, F	0.25	0.32	−0.37 to 0.87	Fair	75	100
10—F, F	0.33	0.24	−0.14 to 0.80	Fair	66.7	100
11—P, P	0.23	0.23	−0.21 to 0.67	Fair	58.3	100
12—P, P	0.33	0.26	−0.17 to 0.84	Fair	66.7	100
13—P, P	0.17	0.23	−0.27 to 0.61	Poor	66.7	91.7
14—F, F	0.09	0.09	−0.09 to 0.27	Poor	41.7	100
15—P, P	0.07	0.20	−0.32 to 0.45	Poor	50	91.7
16—F, F	0.09	0.21	−0.33 to 0.50	Poor	50	91.7
17—P, F	0.35	0.17	0.01 to 0.69	Fair	58.3	100
18—P, P	0	—	—	Poor	33.3	100
19—F, P, P	0.41	0.23	−0.03 to 0.86	Moderate	66.7	100

Note. — = data not calculated.

3.1. IRR Computed by Weighted Kappa

As seen in Table 5, weighted kappas ranged from −0.29 (worse than expected by chance) to 0.54 (moderate agreement) for the 19 candidates. Standard errors and 95% confidence intervals are shown for k_w. The majority of coefficients (n = 11) were in the poor or worse than expected by chance strength of agreement range, n = 6 coefficients were in the fair strength of agreement range, and n = 2 coefficients were in the moderate strength of agreement range; no coefficients fell in the "Good" or "Very Good to Perfect" strength of agreement categories (see Table 2). Fitting a normal distribution to the kappa coefficients resulted in a mean of 0.17 (poor strength of agreement) and a standard deviation of 0.19.

A number of candidates were evaluated by the same evaluators, and we were therefore able to aggregate these results to more precisely estimate IRR. Candidates 12 and 13 had the same evaluators: based on 24 paired ratings, k_w = 0.34 (fair strength of agreement). Candidates 14 and 15 had the same evaluators: based on 24 paired ratings, k_w = 0.38 (fair strength of agreement). Candidates 7, 8, 10, and 11 had the same evaluators: based on 48 paired ratings, k_w = 0.19 (poor strength of agreement). In these comparisons, a more precise estimate of k_w ranged from 0.19 to 0.38, or poor to fair strength of agreement.

3.2. IRR Computed by Percentage Agreement

As seen in Table 5, percentage agreement is shown in two ways, either as an exact match, or as exact plus adjacent agreement, as was carried out in an earlier paper [17]. Again, even though it is considered a mistake to compute IRR using percentage agreement [10], percent agreement has often been reported in the PACT (and now the edTPA [26]) literature. Exact agreement ranged from 33.3% to 75%; in only one instance was IRR accept-

able with greater than 70% agreement [15]. As would be expected, IRR increased in every case when using the methodology of exact plus adjacent agreement [17], which ranged from 83.3% to 100%.

3.3. Qualitative Analysis of Interviews

3.3.1. Confidence, Challenges, and Changes Evaluators Would like to Make to the PACT

Ten PACT evaluators, with a median of 3.75 years of experience, were interviewed for this study. On average, the evaluators took about 2 (SD = 0.6) hours to complete an evaluation, and usually in 1 day (n = 8), but sometimes 2 days (n = 2). The majority found the initial PACT training and annual calibration to be useful, even though several had negative comments about the calibration such as: "didn't enjoy it", "a necessary pain", "could be much better", "I hated them", and "I always struggled".

All of the evaluators felt confident conducting a PACT evaluation, even though the evaluators expressed they would feel more confident under the following circumstances:

> Evaluator 1: It would be fun to sometime have two to three people look at it. I often wonder what someone else would have done with the evidence at hand.

> Evaluator 9: 9 or 10 for confidence, but honestly I would give myself an 8 when I started thinking about how I might stack up with other evaluators.

These quotations express the evaluators' concerns about how their evaluations would compare to others. Perhaps related to this concern is a quote from another evaluator desiring more collaboration, "If I could have collaborated with a colleague I would have been a 10." Other evaluators expressed that they were less confident when faced with tough or borderline cases, and where evidence was missing:

> Evaluator 2: . . . there are tough ones occasionally. Not very confident when they have the 'bones' of good teaching but didn't include everything that was asked for. Question is, if evidence is missing, is it a 1 even if the teaching practice itself is passable?

So, even though all the evaluators expressed confidence in their abilities to evaluate PACTs, several expressed less confidence when they wondered how other evaluators would assess the same evidence. Seemingly related to this was a desire by some evaluators for more collaboration, which would also help evaluators assess tough or borderline cases.

Regarding challenges that the evaluators faced when evaluating PACTs, three mentioned they were unsure about evaluating the evidence (e.g., "Sometimes I am not sure I am interpreting rubrics correctly"). Five evaluators mentioned the difficulty of assessing when there was missing evidence, or missing links. For example, one evaluator stated:

> Evaluator 6: Hardest thing would be grading PACTs that have missing evidence. I had to be sure not to fill in the gaps for them. If the evidence isn't there, then it isn't there.

Other challenges evaluators mentioned were poor videos (mentioned by two evaluators), poor writing (two evaluators), and candidates not following directions (two evaluators).

Therefore, the evaluators expressed uncertainty when evaluating the evidence, including uncertainty with rubric interpretation. Furthermore, several evaluators expressed consternation when there was missing evidence in the PACT submissions, which also came out when the evaluators were interviewed about their confidence. Poor quality in the video presentations and in writing quality further lead to challenges during evaluation, which would also make it more challenging to assess the evidence.

Changes evaluators would like to make to the PACT evaluation process included using the PACT as feedback for the candidates. For example:

> Evaluator 1: How would they [the candidates] learn from it? The feedback is something that could be improved upon.

> Evaluator 9: Use the actual thing to better your practice, not something that has no validity for them or is simply a hoop. Something that they can use in their first year is something that needs to happen eventually for all these TPEs.

The above quotes express a desire that the PACT be used to provide important feedback to the candidates, so that they could benefit from their experience of preparing for and taking the PACT, and then, using this knowledge to improve their teaching practice going forward. Evaluators also expressed a desire for the PACT evaluation process to be more face to face, or community based. For example:

> Evaluator 3: I would say more face to face. Tone should be a community-based thing.
>
> Evaluator 8: I wish there would be a way to meet the candidates and make it more personal. Sit down and discuss comments with the candidate. That would be so great. Face to face is always a good thing and might provide them with an opportunity to respond in real time.

Again, as when the evaluators were asked about their confidence in scoring the PACT, evaluators expressed a desire for more collaboration:

> Evaluator 4: More collaboration! But I know that is hard because of others' schedules.
>
> Evaluator 9: PACT is so isolating for everyone!

The quotes above express the evaluators' belief that the PACT could be improved if it was used to provide specific feedback to the candidates on how to improve their teaching practice. Additionally, it was suggested that the PACT could be improved if it was less isolating and included more face-to-face time with the candidates. It was suggested further that the PACT could be improved by increasing collaboration among the evaluators.

In the following sections, we focus on the actual step-by-step process of evaluating the PACTs, the extent to which the evaluators made use of the Thinking Behind the Rubrics (TBR) document, and how the evaluators addressed the Academic Language (AL) rubrics.

3.3.2. The Step-By-Step Process of Evaluating the PACT

Evaluators were trained to complete evaluations in order, starting with a review of the background information provided by the candidate describing the student population, and proceeding throughout the PACT in order from Planning to Instruction to Assessment to Reflection, while addressing Academic Language throughout.

When the candidates were asked to describe their evaluation process in detail, the prescribed progression for the PACT evaluation process seemed roughly in order for only two evaluators. From the interview responses from two other evaluators, the ordering of the evaluation process was not possible to discern. In contrast, there was evidence of an inconsistent evaluation process for the remaining six evaluators. For example:

> Evaluator 1: I don't adhere to my process the same every time really.
>
> Evaluator 3: I switch it up a lot on how I assess them actually.
>
> Evaluator 6: I watch the video first before planning and then I tackle planning. This might be out of order, but it really helped me to see them first before I read their plans so I can see if there was alignment.

The other three evaluators expressed similar inconsistencies in their evaluation process. Therefore, a majority of the evaluators stated that they followed a variable, rather than a systematic, sequential evaluation process, which is at variance from the PACT training protocol.

3.3.3. Thinking behind the Rubrics (TBR)

PACT evaluators are trained to consistently refer to the TBR document [21], which provides a more in-depth explanation of rubric scores, and even guidance about discernment between adjacent rubric levels. Only one evaluator reported using the TBR document "a lot". One did not use the TBR document the first year, but then tended to use it more. On

the other hand, four evaluators mentioned they did not use the document at all, with one evaluator stating, "I have no idea what that is". One evaluator stated they used the TBR document "very rarely". Three stated they used the TBR document in the beginning, then stopped using it, or only used it for a borderline or tough case. For example, one evaluator said, "I used them a lot in the beginning, and then that waned as the year went on . . . I typically used them for borderline PACTs for sure."

So, even though PACT evaluators were trained to consistently refer to the TBR document during evaluation of the candidates' portfolios—a document that could help the evaluators to assign rubric scores correctly—the majority did not.

3.3.4. Academic Language

Half of the evaluators expressed frustration with evaluating Academic Language (AL) and that it was hard to evaluate. For example:

> Evaluator 3: Certain rubrics are written where a 2 is good and others have a much higher expectation where a 2 seems really low. AL is incredibly rigorous on the spectrum, and somewhat unfair because it is so sophisticated for novice teachers. I am always on high alert for AL. It is the hardest one for me to grade.
>
> Evaluator 6: I hated AL on the PACT. Pretty narrow and annoying to grade.
>
> Evaluator 8: This is the part that I had the most difficulty with. When I trained I remember that everyone had problems with AL. A lot missing with the rubrics and expectations in terms of what teachers should know how to do. Didn't know how to look for AL in the writing...most of the time I gave 2's in AL.

The evaluators expressed strong feelings when they assessed AL and had difficulty in its assessment. Furthermore, two evaluators reported assessing AL out of order, in the beginning, when it should be assessed throughout as per PACT training protocol. For example, one evaluator said, "I addressed them all in the beginning to get them out of the way." Likewise, another evaluator stated, "I would start with the AL rubrics and then do the rest."

In summary, the main findings that emerged from the interviews were (a) that the evaluators were inconsistent in their evaluation process and deviated from scoring protocol; (b) there was a lack of, or inconsistent, use of the "Thinking Behind the Rubrics" document; and (c) Academic Language was the most challenging task to evaluate. The most frequently mentioned challenges for evaluators were candidates missing links or evidence, making the PACT hard to evaluate. Suggested changes included that the PACT should be used more to provide feedback to the candidates, and a desire for a more human connection, including more collaboration (see Appendix A for Interview Questions).

4. Discussion

4.1. IRR Estimates

4.1.1. Weighted Kappa

Weighted kappa (k_w) is an estimate of IRR that takes chance agreement between two raters into account, and also weighted closer ratings higher than more distant ratings (for example, a 1 vs. 2 rating was weighted higher than a 1 vs. 3 rating). Our overall, average estimate of IRR using k_w was 0.17, indicating a poor strength of agreement. Because a number of candidates were evaluated by the same evaluators, we were able to compute more precise k_w estimates of IRR of 0.34 (for two candidates), 0.38 (for two candidates), and 0.19 (for four candidates). Prior studies that also report more precise estimates of IRR in TPAs also find low IRR (e.g., for intra-class correlations in the CalTPA [16]). In the only other studies that reported kappa for IRR in the PACT [19,20] (using the same data), their polytomous kappa, most similar to our k_w, was 0.33 (fair strength of magnitude), which was a figure we found in a subset of our analyses. In general, these more rigorous ways of computing IRR found poor to fair strength of agreement, which is contrary to the

early reports of what PACT researchers reported [17], and has often been repeated in the literature (e.g., [6]).

4.1.2. Percentage Agreement

Our findings of the consensus estimates of IRR being low when using k_w are in contrast to prior studies that reported estimates of IRR as being very good when using percentage agreement (e.g., [17]). Our data demonstrated the inflation in IRR that results when percentage agreement is computed as exact plus adjacent agreement. For example, our data showed that overall percentage agreement went from 55.7% (for exact agreement between two raters) to almost 100% when percentage agreement included adjacent agreement (see Tables 4 and 5). Even though percentage agreement is an inferior measure of IRR [10], we computed it to compare to prior findings [17,19,20]. For example, Porter and Jelinek [20] reported exact agreement between two evaluators as 66%, compared to our finding of 56%. All of these data were below the suggested 70% or greater consensus estimate for reliable IRR [15], and importantly, did not adjust for chance agreement.

We were able to compare our computation for exact agreement using our PACT data in the present study to the most recent edTPA report [26]. Our results, 55.7% exact agreement, are close to the results (56.7%) reported in the edTPA report (p. 9). Our results for exact plus adjacent agreement (96.9%) are similar to the overall result reported in the SCALE [26] (p. 9) report (95.9%). We are concerned that the edTPA [26] continues to report percentage agreement, despite it being "definitively rejected as an adequate measure of IRR" [10], (p. 4).

4.1.3. Kappa N

SCALE [26] also computes IRR as kappa *n* (k_n), which seems to be inappropriately used (see [5,27]). Brennan and Prediger [27] stated: "if nonfunctional categories are used in a study, k_n may be artificially large" (p. 693). Scott [28] stated that "the index is based on the assumption that all categories in the dimension have equal probability of use . . . by both coders. This is an unwarranted assumption for most behavioral and attitudinal research. Even though *k* categories may be available to the observers, the phenomena being coded are likely to be distributed unevenly, and in many cases will cluster heavily in only two or three of them" (pp. 322–323). As observed in the PACT data (see Table 3), a score of 4 is rarely given. In the edTPA data, scores of 1 and 5 are rarely given [5]. When we computed our PACT data as kappa *n*, IRR increased from $k_w = 0.17$ to $k_n = 0.41$, a result comparable to the kappa n (0.46) was reported for the edTPA data [26] (p. 9). Differences in our k_n estimate and that from the edTPA could partly be explained by the PACT using 4 rating categories vs. the edTPA using 5 rating categories (more categories lead to larger k_n estimates). Gitomer et al. [5] reported that the IRR data reported by edTPA are "implausibly high" (p.15).

4.1.4. Measurement Error in IRR Estimates

In our data, 3 of the 19 candidates initially passed the PACT, but upon double scoring, were found to fail. The initial passing PACT status for these candidates was preserved. It is an open question whether these candidates should also undergo rating by a third evaluator as a tiebreaker, as is carried out for those who initially fail the PACT.

The low IRR we observed on the PACT suggests that the observed ratings contain a large amount of measurement error [10] (p. 5). Possible reasons for the low IRR we observed could be because of restriction of range of the rating scale (1 to 4, with 4s rarely given), or perhaps difficulty in observing or quantifying the variables of interest [10] (p. 5), which was partly corroborated by the data we collected in the interviews. For example, interviewees stated they had difficulty rating the rubrics in Academic Language (see Duckor et al. [29] who also reported challenges in assessing Academic Language). Assuming the evaluators were properly trained, it is possible that once trained, the evaluators drifted from the criterion performance. This is possible, in that we found evidence that evaluators strayed

from training protocol, and did not regularly make use of the evaluation aid, "Thinking Behind the Rubrics." It is likely a combination of these observations that led to the consensus estimate of low IRR we observed in our study.

4.2. Cognitive Task Analysis

The interviews and qualitative data reveal two overarching themes with respect to the process of evaluating a teacher performance assessment (TPA) such as the PACT. First, in reviewing and scoring a TPA, evaluators engage in highly complex cognitive problem solving. There are many variables that must be considered simultaneously, such as use of academic language, evidence of planning, and evidence of implementation by identifying lesson artifacts. We now know that the limitations of our working memory resources require us to "swap out" these variables under consideration [30]. During this mostly unconscious process of managing our mental resources, there arises a potential of losing track of variables that have been considered and those that have not, resulting in errors of omission, as well as incorrectly classified instances of the variables, resulting in errors of commission.

Second, as the interviews and qualitative data demonstrate, evaluators, for the most part, work in siloed environments while assessing a TPA. The desire for additional collaboration, expressed by several evaluators, shows the potential benefit of having "another set of eyes" during the evaluation process. Having a closer collaboration during assessment would enable the two evaluators to "negotiate" what they are reviewing against the rubric and with each other as a constant check on the accuracy and consistency of their scoring. This type of collaboration, albeit beneficial for IRR, lacks practicality in having two evaluators in the same location at the same time. However, advances in technology are now able to mitigate this limitation by enabling both synchronous and asynchronous evaluation by two evaluators.

One possible solution to both the issue of complexity and collaboration is the use of a proxy for the second evaluator in the form of a standard protocol based on measurable expertise. The protocol would need to be generated from multiple expert evaluators who have achieved a very high degree of IRR when comparing the results of their teacher performance assessments. Such a protocol assumes that we are able to capture the unobservable thoughts, decisions, and judgments individual expert evaluators make when categorizing an artifact as meeting specific criteria and assigning a corresponding score.

Methods to capture this expertise exist in the form of Cognitive Task Analysis (CTA). CTA employs semi-structured interviews with multiple subject matter experts (SMEs) to capture the knowledge and skills they use while solving difficult problems and performing complex tasks [31]. Capturing unobservable thoughts, decisions, and judgments from multiple experts is required as studies show that experts may omit up to 70% of the critical information when describing how to replicate expert performance to others [32]. When three to four experts are used to aggregate their knowledge and skills as a gold standard protocol, this "70% rule" can be reversed. Meta-analytic research reveals significant performance improvement when the results of CTA are used for instruction and as job aides while executing complex and difficult tasks [33]. It would appear reasonable, then, to expect that the accuracy and IRR of teacher performance assessment would be increased in relation to the fidelity of following the gold standard protocol.

4.3. Study Limitations

Four study limitations were identified. First, a limitation of the current study is its small sample size. This was a pilot study, and could be improved by a larger sample size study, which more precisely quantifies IRR. Second, prior to the interviews being conducted, the interviewer knew IRR was low and also knew the evaluators, introducing the possibility of confirmation bias [34]. This possibility, however, was somewhat mitigated by having two researchers independently analyze the interview data. Third, limitations of Cohen's kappa such as prevalence or bias problems [10] were examined. When the

marginals for the 4 × 4 double-scored coding matrices were examined for each candidate (see Table 3 for source), no systematic bias problems were observed. We did, however, observe prevalence problems due to the fact that ratings of 4 were rarely given (<2%). This may have resulted in an unrepresentatively lower IRR estimate [10] (p. 6). Fourth, the same evaluators did not rate all 19 candidates, so we are only doing pairwise comparisons: for each weighted kappa, we have 12 possible comparison points and the 95% CIs reported in Table 5 are only for these pairwise comparisons. We cannot aggregate these points along multiple evaluators to compute a 95% CI for a kappa with higher *n*, except for a few candidates who had the same evaluators (see Section 4.1.1 Weighted Kappa). Because all candidates were not rated by the same evaluators, our design was not fully crossed. "Fully crossed designs . . . allow for systematic bias between [evaluators] to be assessed and controlled for in an IRR estimate, which can improve overall IRR estimates" [10] (p. 3).

4.4. Summary and Suggestions

Even though the PACT has been discontinued as a TPA at our institution, it has evolved into the edTPA, which is used nationally at several institutions, including our own. We sought to examine IRR in our candidates' edTPA data, but when we requested double-scored data, we were not given access because the data are proprietary to Pearson Inc. Because we cannot access the double-scored data behind the scores for our edTPA candidates, it is impossible to assess IRR using the methods we used in this study (weighted kappa).

In the present study, we have demonstrated mostly poor consensus estimates of IRR in the PACT in our sample. As a solution, IRR could possibly be increased if cognitive task analysis (CTA) were implemented in the evaluation process. Going forward, it is important that assessing IRR in high stakes teacher performance assessments such as the edTPA is transparent [5] and needs to be carried out on an ongoing basis. Ultimately, increasing the IRR of performance assessments for pre-service teacher candidates could enhance society's best efforts at putting the best teachers into the classrooms.

We ask that SCALE makes anonymized double-scored data from the edTPA available so that researchers in the teaching profession can perform independent analyses. If it is found that edTPA data are not reliable, perhaps the edTPA could be discontinued as a high stakes summative assessment (see Gitomer et al. [5], who suggest a moratorium using the edTPA). If changes are made to increase the IRR of edTPA to acceptable levels, then perhaps it could be continued as a summative assessment.

Under the current climate of increased pressure on schools of education to be held accountable for producing high quality teacher graduates, it is only fair that performance assessments be held equally accountable. Accountability includes being transparent so that assessment results can be publicly scrutinized. In such a way, efforts to train and certify competent teachers may become more successful.

Author Contributions: Conceptualization, S.A.L., K.P. and K.Y.; methodology, S.A.L., K.P. and K.Y.; validation, S.A.L.; formal analysis, S.A.L.; investigation, S.A.L. and K.P.; data curation, S.A.L. and K.P.; writing – original draft preparation, S.A.L., K.P. and K.Y.; writing – review & editing, S.A.L., K.P. and K.Y.; visualization, S.A.L.; funding acquisition, S.A.L. and K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by a grant from the "California Council on Teacher Education (CCTE) 2014-2015 Quest for Teacher Education Research," which was provided by a State Chapter Support Grant from the American Association of Colleges for Teacher Education (AACTE).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of the University of Southern California (protocol code UP-15-00075 on 2/6/2015).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw quantitative data from which estimates of inter-rater reliability were computed are shown in Table 3 of the current study. Qualitative data are openly available from a publicly archived dataset and can be accessed at: https://www.researchgate.net/publication/355316717_IRR_Qualitative_DB_for_RG_data_depository_EduSci, accessed on 12 August 2021. DOI:10.13140/RG.2.2.27423.48808.

Acknowledgments: We thank Rand Wilcox for statistical consultation, an anonymous statistical reviewer, and Judith Mitchell for her suggestions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Interview Questions

The evaluators were asked the following follow-up questions.

1. How many years have you been a PACT evaluator?
    a. What is your content area expertise?
    b. What is your occupation outside of being a PACT evaluator?
    c. Do you evaluate any other TPAs for any other institution?
        i. If so, is flipping between the two TPAs challenging for you?
2. How long has it been since your initial PACT training?
    a. How useful did you find this training?
    b. How useful is the yearly calibration?
3. On average, how long does it take you to complete a PACT evaluation?
4. For PACTs that are double scored, do you know if they initially passed or failed?
5. Please take me through your process of evaluating PACTs. Typically, what do you do first, second, etc., all the way through to the completion of the PACT evaluation?
6. How do you approach the comments portion of the PACT evaluations, and how important do you feel this feedback is to the candidate?
7. How confident do you feel in conducting a PACT evaluation?
    a. Do you feel you received sufficient training? Do you have any feedback about your training experience?
8. To what extent do you use supporting documents such as "Thinking Behind the Rubrics" when evaluating PACTs?
9. Do you normally complete PACTs in one day, or does it take you multiple days?
    a. If multiple, how many days?
10. Do you get fatigued when you evaluate PACTs, and what steps (if any) do you take to reduce this fatigue?
    a. Do you feel you are as fresh on the latter rubrics as you are on the earlier rubrics?
11. How do you address the Academic Language rubrics?
12. What challenges, if any, do you encounter when evaluating PACTs?
13. What changes, if any, would you make to the PACT evaluation process?

References

1. Boshuizen, H.P.A. Teaching as regulation and dealing with complexity. *Instr. Sci.* **2016**, *44*, 311–314. [CrossRef]
2. Reagan, E.M.; Schram, T.; McCurdy, K.; Chang, T.; Evans, C.M. Politics of policy: Assessing the implementation, impact, and evolution of the performance assessment for California teachers (PACT) and edTPA. *Educ. Policy Anal. Arch.* **2016**, *24*, 1–22. [CrossRef]
3. Lalley, J.P. Reliability and validity of edTPA. In *Teacher Performance Assessment and Accountability Reforms: The Impacts of Edtpa on Teaching and Schools*; Carter, J.H., Lochte, H.A., Eds.; Palgrave, MacMillan: New York, NY, USA, 2017; pp. 47–78.

4. Hebert, C. What do we really know about the edTPA? Research, PACT, and packaging a local teacher performance assessment for national use. *Educ. Forum* **2017**, *81*, 68–82. [CrossRef]
5. Gitomer, D.H.; Martinez, J.F.; Battey, D.; Hyland, N.E. Assessing the assessment: Evidence of reliability and validity in the edTPA. *Am. Educ. Res. J.* **2019**, *58*, 3–31. [CrossRef]
6. Okhremtchouk, I.; Seiki, S.; Gilliland, B.; Ateh, C.; Wallace, M.; Kato, A. Voices of pre-service teachers: Perspective on the performance assessment for California teachers (PACT). *Issues Teach. Educ.* **2009**, *18*, 39–62.
7. Bird, J.; Charteris, J. Teacher performance assessments in the early childhood sector: Wicked problems of regulation. *Asia-Pac. J. Teach. Educ.* **2020**, 1–14. [CrossRef]
8. Charteris, J. Teaching performance assessments in the USA and Australia: Implications of the "bar exam for the profession". *Int. J. Comp. Educ. Dev.* **2019**, *21*, 237–250. [CrossRef]
9. Stacey, M.; Talbot, D.; Buchanan, J.; Mayer, D. The development of an Australian teacher performance assessment: Lessons from the international literature. *Asia-Pac. J. Teach. Educ.* **2020**, *48*, 508–519. [CrossRef]
10. Hallgren, K.A. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor. Quant. Methods Psychol.* **2012**, *8*, 23–34. Available Online: https://pubmed.ncbi.nlm.nih.gov/22833776/ (accessed on 12 August 2021) [CrossRef]
11. McClellan, C.A. Constructed-response scoring—Doing it right. *R D Connect.* **2010**, *13*, 1–7.
12. McMillan, J.H. *Classroom Assessment. Principles and Practice for Effective Standards-Based Instruction*, 5th ed.; Pearson: Boston, MA, USA, 2010.
13. Pufpaff, L.A.; Clarke, L.; Jones, R.E. The effects of rater training on inter-rater agreement. *Mid-West. Educ. Res.* **2015**, *27*, 117–141.
14. Sherman, E.M.S.; Brooks, B.L.; Iverson, G.L.; Slick, D.J.; Strauss, E. Reliability and validity in neuropsychology. In *The Little Black Book of Neuropsychology: A Syndrome-Based Approach*; Schoenberg, M.R., Scott, J.G., Eds.; Springer: Boston, MA, USA, 2011; pp. 873–892. [CrossRef]
15. Stemler, S.E. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Pract. Assess. Res. Eval.* **2004**, *9*, 1–11. [CrossRef]
16. Riggs, M.L.; Verdi, M.P.; Arlin, P.K. A local evaluation of the reliability, validity, and procedural adequacy of the teacher performance assessment exam for teaching credential candidates. *Issues Teach. Educ.* **2009**, *18*, 13–38.
17. Pecheone, R.L.; Chung, R.R. Evidence in teacher education: The performance assessment for California teachers (PACT). *J. Teach. Educ.* **2006**, *57*, 22–36. [CrossRef]
18. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
19. Porter, J.M. Performance Assessment for California Teachers (PACT): An Evaluation of Inter-Rater Reliability. Ph.D. Thesis, University of California, Davis, CA, USA, 2010.
20. Porter, J.M.; Jelinek, D. Evaluating inter-rater reliability of a national assessment model for teacher performance. *Int. J. Educ. Policies* **2011**, *5*, 74–87.
21. Thinking Behind Rubrics. Available online: https://www.uwsp.edu/education/documents/edTPA/Resource10.doc (accessed on 29 September 2021).
22. GraphPad QuickCalcs. Available online: http://www.graphpad.com/quickcalcs (accessed on 12 August 2021).
23. Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220. [CrossRef]
24. Kappa as a Measure of Concordance in Categorical Sorting. Available online: www.vassarstats.net/kappa.html (accessed on 12 August 2021).
25. *ATLAS.ti (Version 8.4) [Qualitative Data Analysis Software]*; ATLAS.ti Scientific Software Development GmbH: Berlin, Germany, 2019.
26. Stanford Center for Assessment, Learning and Equity (SCALE). *Educative Assessment and Meaningful Support: 2019 edTPA Administrative Report*; Stanford Center for Assessment, Learning and Equity (SCALE): Palo Alto, CA, USA, 2021.
27. Brennan, R.L.; Prediger, D.J. Coefficient kappa: Some uses, misuses, and alternatives. *Educ. Psychol. Meas.* **1981**, *41*, 687–699. [CrossRef]
28. Scott, W.A. Reliability of content analysis: The case of nominal scale coding. *Public Opin. Q.* **1955**, *19*, 321–325. [CrossRef]
29. Duckor, B.; Castellano, K.E.; Tellez, K.; Wihardini, D.; Wilson, M. Examining the internal structure evidence for the performance assessment for California teachers: A validation study of the elementary literacy teaching event for tier 1 teacher licensure. *J. Teach. Educ.* **2014**, *65*, 402–420. [CrossRef]
30. Ma, W.J.; Husain, M.; Bays, P.M. Changing concepts of working memory. *Nat. Neurosci.* **2014**, *17*, 347–356. [CrossRef] [PubMed]
31. Clark, R.E.; Feldon, D.F.; van Merrienboer, J.J.G.; Yates, K.A.; Early, S. Cognitive task analysis. In *Handbook of Research on Educational Communications and Technology*, 3rd ed.; Spector, J.M., Merrill, M.D., van Merrienboer, J., Driscoll, M.P., Eds.; Lawrence Erlbaum Associates: New York, NY, USA, 2008; pp. 577–593. [CrossRef]
32. Feldon, D.F.; Clark, R.E. Instructional implications of cognitive task analysis as a method for improving the accuracy of experts' self-reports. In *Avoiding Simplicity, Confronting Complexity: Advances in Studying and Designing Powerful (Computer-Based) Learning Environments*; Clarebout, G., Elen, J., Eds.; Sense Publishers: Rotterdam, The Netherlands, 2006; pp. 119–126. [CrossRef]

33. Tofel-Grehl, C.; Feldon, D.F. Cognitive task analysis-based training: A meta-analysis of studies. *J. Cogn. Eng. Decis. Mak.* **2013**, *7*, 293–304. [CrossRef]
34. Oswald, M.E.; Grosjean, S. Confirmation bias. In *Cognitive Illusions. A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*; Pohl, R.F., Ed.; Psychology Press: Hove, UK, 2004; pp. 79–96.