Educational Research for Policy and Practice (2023) 22:461–478
https://doi.org/10.1007/s10671-023-09340-9

ORIGINAL ARTICLE

From face-to-face to remote evaluation of teacher-education candidates during the COVID-19 pandemic

Judy Goldenberg[1,2,3] · **Doron Niv**[2]

Received: 9 November 2021 / Accepted: 21 March 2023 / Published online: 20 May 2023

Abstract

Although teacher's personality is an essential component of successful classroom learning, most teacher education programs accept students solely on the basis of scholastic ability scores such as school grades, national test scores (SAT, GRE) or undergraduate college transcripts. To ensure suitability to teaching, a personality-evaluation system was developed in Israel for teacher education candidates. This evaluation system includes non-cognitive measures, such as group dynamic exercises, simulations, a teaching exercise, situational judgement tests, personality tests and an inter-personal interview, all performed face-to-face (FTF) at a testing center. The outbreak of COVID-19 in 2020 brought about lockdowns and social distancing, precluding the administration of this FTF evaluation system. Therefore, the development team adapted the system to online remote testing, using Zoom technology. The present study examined the effect of this transition to remote evaluation on the quality of selection for teaching, looking at suitability-to-teaching scores and the subjective views of applicants and evaluators. A comparison of the 2020 remote scores with the 2019 FTF scores revealed that scores on remote evaluation were slightly lower than FTF scores, and were more centralized. While the candidates found that remote evaluation provided fewer opportunities to express themselves, both candidates and evaluators were satisfied with the administration and convenience of the evaluation day. The Discussion chapter summarizes the unique affordances and constraints of remote evaluations and presents suggestions for changes which might be made when moving an assessment online that could take advantage of this new environment.

Keywords COVID-19 pandemic · Remote testing · Teacher-education studies · MESILA

✉ Judy Goldenberg
Judyg@zahav.net.il

Doron Niv
doronn@macam.ac.il

[1] Talpiot Academic College of Education, Holon, Israel

[2] The Mofet Institute, Tel Aviv-Yafo, Israel

[3] The Mofet Institute, Tel Aviv, Israel

1 Introduction: the need for personality-based selection for teacher-education programs

Nothing is more important to education than the individual teacher. All other factors—educational programs, physical environment, class size, and budget, important as they are, are secondary. It is the teacher as an educator and instructor who has the strongest impact on children's educational experience thus it is necessary to invest in high-quality educator preparation (Darling-Hammond & Hyler, 2020). Setting clear screening processes and criteria for admission to teacher-education institutions to help select candidates who will succeed as good classroom educators is a crucial step, which has, to date, received little research attention (Klassen & Kim, 2019).

Alongside the traditional entrance measures, which relied on matriculation scores and entrance exams, there is growing recognition that teacher-education schools' evaluation should also assess the personal and professional suitability of candidates (Casey & Childs, 2011; Klassen & Kim, 2019). One such system developed to evaluate the personality of the teacher education candidates in Israel, is the innovative MESILA (Hebrew acronym for Unique Student Evaluation for Teacher-education Studies). This system evaluates candidates' tendencies, behaviors, values, motivations, expectations, and interpersonal abilities. This full-day evaluation system includes non-cognitive alternative prediction measures, such as interpersonal group exercises and personality measures which will be described in the following chapter, all performed face-to-face (FTF) at a testing center.

However, the outbreak of COVID-19 in 2020, and the ensuing social distancing, lockdowns, and quarantines, precluded FTF evaluations and group interaction. In order to continue the evaluation of applicants for teacher-education studies, the selection system was changed to remote evaluations, using Zoom technology.

The research objective of this article was to measure the effect of the transition to remote selection, comparing the two modes of evaluation: remote evaluation scores during COVID-19 to that of FTF evaluation conducted in the previous year. Three research questions guided our research:

1. What quantitative differences were found when comparing data of the remote evaluation during COVID-19 to that of FTF evaluation conducted in the previous year?
2. What were the subjective perceptions of the candidates to the remote evaluation as compared to perceptions of candidates who underwent FTF evaluations during the previous year?
3. What were the subjective perceptions of the evaluators who participated in remote evaluation as compared to evaluator perceptions using FTF evaluations during the previous year?

To examine these research questions, we have organized this article in the following manner: first, the background chapter presents the need for personality assessment to evaluate candidates for teacher education, and illustrates how this evaluation was performed face-to-face using MESILA. Next we present the difficulties in evaluation resulting from the COVID-19 pandemic, and describe the transition to remote evaluations. The literature chapter presents an overview of findings on the influence of remote evaluations on candidates and on evaluators. The following chapters present the research method and the results for the three research questions. A discussion of these results sheds light on the affordances of remote evaluations, presents suggestions for changes which might be made

when moving an assessment online and possible contributions to evaluating the suitability of teacher candidates to the growing needs of digital literacy and online teaching.

2 Background: the need to evaluate candidate's personality for teacher education

Teachers are the most important and significant factor in educating children, and in many countries, the education system is one of the largest government employers. To ensure good classroom educators, careful selection of teacher education candidates is vital (Klassen & Kim, 2019). Reviews of screening processes for teacher-education schools showed the overwhelming use of cognitive measures such as grades and achievements (e.g., school grades, national test scores such as SATs or GREs) (Mihelic et al., 2018; Roloff et al., 2020). However, these commonly used evaluations are based solely on scholastic abilities, and do not measure personality variables related to teacher success.

The need to evaluate personality variables among teacher-education candidates has introduced into the field of teacher education measures frequently used in occupational psychology to evaluate candidates in the work force. Schmidt (2016) surveyed prevalent methods of evaluating candidates in order of their validity in predicting job success and their incremental validity beyond mental ability tests. The measures commonly used were work samples, structured interviews, peer evaluations, biographical questionnaires, integrity tests, unstructured interviews, group dynamics, work ethic tests (such as "conscientiousness"), and recommendations. Based on this list of potential measurement tools, as well as past experience of the research team creating selection batteries, a personality based evaluation system for teacher-education studies was developed, as will be described in the following chapter.

3 Developing an evaluation system for teacher-education studies candidates

3.1 FTF teacher education evaluation

With the growing recognition of the need for a screening system that addresses the personal and professional suitability of candidates for teacher-education studies, the MESILA FTF evaluation was developed in Israel as a national test.

MESILA is a full-day evaluation, during which professionals evaluate the applicants using standard tools which include: interactive group dynamic situations, a teaching exercise, simulations of behaviors required of teachers, interpersonal interviews and peer evaluations. In addition, the candidates complete computerized biographical questionnaires, a situational judgement test and computerized personality tests (Goldenberg, 2018, 2020). In the FTF mode, these evaluations are conducted in group settings, with several evaluators observing the candidates, grading their suitability on forms especially designed for each exercise. Following each task, the evaluators complete a rating scale on which they rate several variables relevant to the individual exercise. For example, after the teaching exercise the rating includes variables related to organization of the material and verbal expression. After completing the interpersonal simulation, they rate variables such as empathy and assertiveness. Thus, each exercise generates an

independent test score. A final MESILA score on suitability to teacher-education studies is computed consisting of a weighted composite score of all of the exercises.

At the end of the evaluation day, a report is issued for each candidate, with sub-scores for the main measures, a final score of suitability to teaching, and a short, written appraisal of the candidate's strengths and areas that need improvement. Scores on these non-cognitive measures, in conjunction with matriculation and psychometric scores, present a full, rounded portrait of the candidates. Research has shown that candidates' scores on MESILA were found to have predictive validity for various criteria of success in academic teacher-education studies, for pre-service training ratings, for classroom teaching evaluations and for measures such as receiving tenure as teachers (Goldenberg, 2020).

3.2 The COVID-19 pandemic and the development of the remote mode of MESILA evaluation

The outbreak of the COVID-19 pandemic in the beginning of 2020, brought about regulations regarding social distancing in closed areas and inter-city travel, as did alerts about an overall lockdown and quarantines for people infected.

In the world of employment, recruitment and evaluation procedures changed drastically, often being postponed or slowed down in an unprecedented manner, requiring adaptation and using remote evaluation. One notable implication was a switch to synchronous interviews using video conference calls or the telephone (Manella, 2020). This necessitated a change in the MESILA evaluation of student candidates at the very time that applications for higher-education were to be submitted. It was necessary to immediately transform the MESILA FTF evaluations into a system which could be administered remotely, testing each candidate individually, usually from their home. The development team of MESILA adapted the existing FTF tests which were described above into a remote mode of testing using the Zoom technology available on home computers.

The remote MESILA evaluation was conducted through two Zoom meetings: in the first meeting, candidates were given the same computerized tests as previously (the on-line biographical questionnaire, a situational judgement test and personality tests) however due to social distancing they were completed from the candidates' home and not at the central testing center. The second Zoom meeting was held with two evaluators who conducted an in-depth personal interview and performed several exercises with the candidate. In the teaching exercise, the candidate was required to teach a short lesson online with the evaluators acting as pupils in the classroom. In the next stage the candidate was engaged in on-line simulations of situations prevalent in the life of a teacher, with the two evaluators playing either the role of pupils, parents, the principle or peer teachers. There was no interaction with other candidates, and the entire procedure involved the one candidate and two evaluators on line. The instructions for the teaching exercise, simulations, and the interview were similar to those for FTF and remote testing, except for changes necessitated by the test mode.

The main difference between the content of the two testing modes was the cancelation of the group dynamics exercise and the peer evaluations, which were a part of the FTF testing system but were reduced for logistic reasons. At the end of the Zoom meeting, the evaluators graded candidate's suitability to teacher-education studies, rating the same scores as they had on FTF evaluations.

4 Literature review: findings on remote testing

Although research has been published on the effects of telephone- or video-conference interviews (Basch et al., 2020), findings on the relationship between FTF evaluation and remote evaluation is in its infancy. There are indications for the equivalence of the scores in each system, but few studies have examined the validity criterion of remote evaluation tools (Woods et al., 2020). Straus et al. (2001) found that interviewees reported greater difficulty to produce evaluation information in a video-conference interview than in a FTF interview. They also found that in the remote interviews, the range of scores was smaller, and that there were fewer extreme values. The researchers attributed these findings to the interviewers' lack of confidence in the evaluation and unwillingness to reject a candidate based on information received in the distance evaluation, leading to concerns that remote evaluation tools can harm the validity of the evaluation. This concern has been supported in an analysis of evaluators' questionnaires in remote evaluation; the evaluators reported some difficulties with evaluation because they lacked cues from the candidates' non-verbal behavior and body language. However, as the evaluators gained more experience, and had more training in remote evaluation, this difficulty diminished (Helding, 2020).

Researchers have found that converting evaluation tools that require interpersonal communication to computer-mediated tools is more complicated than converting remote cognitive tests and closed questionnaires (McCarthy et al., 2017; Potosky & Bobko, 2004). In their Media Richness theory, Daft and Lengel (1986), claimed that media differ from each other in the richness of information they can process and transmit. FTF communication is the richest, as it allows us to observe many non-verbal cues such as body language and facial expressions. Video-based communication was ranked lower because of the difficulty in discerning non-verbal cues. Thus, we must examine the fidelity of remote tests, compare it to information received through in-person, FTF evaluation, and examine the effect of the evaluation method on the scores achieved.

We will now turn to noting the characteristics of remote evaluation and their possible effects on the evaluation, based on existing professional knowledge. Russell (2015) coined the term *screen relations* to describe the unique characteristics of psychological therapy during a video call. Similarly, many characteristics of remote evaluation require unique understanding and treatment.

4.1 Situational characteristics

Remote testing reduces the evaluator's control over the test conditions, as the candidates can log on to the program in any location, without being monitored, undermining standardization and potentially hindering equity and equal opportunity. There is an increased dependence on technology. Remote evaluation requires the presence of video-call technology and the means to operate it. The three main components of this technology are software, hardware, and internet. As for software, not all candidates have equal mastery of Zoom, although these gaps may have narrowed over time. Hardware is an issue because not all homes are equipped with computers that have the hardware to conduct video calls (camera, microphone, and speakers) that are available during the time scheduled for the test. The issue of internet infrastructure carries its own challenges, as the infrastructure must be able to handle two-way video and audio transmission, with minimum delay. Dependence on technology can increase inequity between groups of different socioeconomic levels

(Vadtal & Georgi, 2020). One way to address this issue was to place computers in accessible places (such as schools or human-resources offices) and invite interested candidates to take the remote admission tests individually, once lockdown was removed.

An additional characteristic is the lack of social presence. In remote evaluation, candidates are on their own, devoid of the presence of other candidates. According to Short et al. (1976), in social situations a higher "perceived presence" of others increases the chances for better, more effective communication, and a sense of lack of presence can lead to communication difficulties among participants. Basch et al. (2020) used a remote video interview to examine the issue of perceived presence, and found that a lower sense of the presence of others was associated with lower scores for the candidates. Evidence was also found that lower perceived presence leads to less use of impression-management techniques and less eye contact, possibly explaining the findings (Fullwood & Finn, 2010; Helding, 2020).

4.2 Psychological characteristics

One psychological effect of remote testing results from missing information due to on-line testing. Social interaction that has a richness of cues (verbal and non-verbal) will increase the evaluation's level of accuracy (Ekman, 2004; Potosky, 2008). Weinberg and Rolnick (2020) noted that when working online, the interviewer's visual attention is focused on the candidate's torso and head, in a "bodiless environment." This limits the ability to register such characteristics as position, gestures, and hand movements, so that ability to deduce the situation is limited. *Gaze awareness,* the ability to absorb and interpret the characteristics of the gaze of others, is also compromised (Slovák, 2007). In addition, FTF evaluation included interpersonal interaction even during downtime such as breaks, when the candidates are not being evaluated formally (and may therefore feel freer).

The characteristics of video calls could create a different discourse dynamic among the speakers who experience delays between the time one person speaks and the next one hears and responds (Roberts & Francis, 2013). The lag time between statement and response generates dynamics of "speaking in turns," which decreases spontaneous interaction (Wegge, 2006), and is a source of frustration (Helding, 2020). These delays do not allow people to interrupt each other, which they can do in a FTF situation, providing the evaluators with important interpersonal information (Sklar, 2020), perhaps making more volatile candidates seem more tolerant.

4.3 Emotional aspects

Participants in online conversation tend to feel less comfortable, less protected, and less emotionally connected to the situation (Holmes & Kozlowski, 2015). Those who make frequent eye contact are perceived by others as more attentive, friendlier, more cooperative, and as giving a sense of security (Slovák, 2007). In a conversation carried out over a computer, when we look into the eyes of the person on the screen, we must look away from the camera (which is usually above the screen), and could therefore be perceived of as *not* looking straight (Wolf, 2020). This type of dynamics is a hindrance to creating a stable interpersonal experience and could contribute to participants feeling confused and bewildered. Additionally, Zoom (and similar programs) allows participants to see themselves at any moment (much like looking at yourself in a mirror) which could increase

social-desirability behaviors and use of impression-management techniques (Helding, 2020; Horn & Behrend, 2017).

4.4 Absence of the group dimension

It is difficult to evaluate interpersonal characteristics by remote testing. The definition of the "good teacher," which underlies the MESILA tests, includes an interpersonal and social characteristic, tested by such measures as ability to collaborate with other team members and be a useful team member, be considerate, able to compromise, have good people skills, and capable of providing sensitive feedback without raising antagonistic feelings. Like other characteristics assessed for teacher-education studies, these too are assessed by different tools—some declarative, where the candidates report on their abilities, and others behavioral, where evaluation is based on observation. FTF testing provided information on these attributes based on interactive group exercises which are missing in remote testing.

Remote evaluation systems which include group exercises have their drawbacks. Schlapobersky (1993), attempting to understand group unconsciousness, did not ask what the person says, but to whom that person speaks. In the Zoom meeting, the addressee of the participants' responses cannot be discerned, so that group phenomena such as coalitions, sub-groups, and dyads could be missed, distorted, or confused. Also, during a remote conversation the ability to tell who is looking directly at other people is compromised (Slovák, 2007).

4.5 Summary of FTF and remote evaluations

The MESILA transition to remote evaluation for teacher-education studies affected the following interfaces: foregoing the group situational exercise, which led to losing important interpersonal information, this in addition to not having peer evaluation; absence of additional participants and observers during the exercises (which increase stress and provide information on behavior under pressure), and lack of information regarding the degree to which mutual feedback was accompanied by empathy.

At the same time, despite the transition to individual evaluation, there were remaining interfaces where the interpersonal dimension could be assessed, and social characteristics discerned when the evaluators participated in the exercises, such as during simulations. The evaluators were instructed to be active and alert to these dimensions during those exercises, as part of the role play. Additionally, the declarative interfaces in the personality questionnaires and personal interview provided important information even in the remote evaluation.

5 Research objectives

The present study was designed to examine the effect of the transition of the MESILA evaluation system for candidates for teacher-education studies from face-to-face to remote evaluation, examining the quality of the evaluation and the participants' perception of the evaluation. Data of the remote evaluation were analyzed and compared to those of FTF evaluation conducted in the previous year. A two-level comparison was performed: first, the candidates' scores in the remote evaluation were examined quantitatively, and compared to scores from the previous year's FTF evaluation. Next, an examination of the

participants' subjective perception of the remote evaluation was conducted using a closed end survey, learning about the perceptions of the evaluators and the candidates.

6 Method

6.1 Population

FTF sample: 860 candidates who underwent FTF evaluation for undergraduate teacher-education studies in 2019.

Remote sample: 908 candidates who underwent remote evaluation for undergraduate teacher-education studies in 2020.

Candidates' demographics were similar in both populations with 74% of the FTF 2019 evaluation candidates women, and 72% in the 2020 remote evaluation. There were no changes regarding the populations, nor were there changes in the content of the evaluation tools, except for those mandated by the transition to remote testing and described above.

A feedback questionnaire was distributed to the participants to rate their satisfaction with the evaluation system. The FTF candidates completed this questionnaire at the test site, however among the remotely tested candidates, the questionnaire was sent by email and the response rate was 28.4% (258 of the 908 who underwent remote evaluation). This rate is consistent with that cited in the literature for e-mail surveys (Yun & Trumbo, 2000), especially those using unverifiable e-mail addresses where the number of undeliverable questionnaires cannot be subtracted from the initial sample to compute the exact response rate (Fincham, 2008). Among the evaluators, almost all (14) responded to the feedback questionnaire. Many of them had participated in both the FTF and remote evaluation, so that their comparison was based on personal information.

6.2 Tools

Three research tools were used: the candidates' scores on the MESILA tests from FTF and remote evaluations, candidates' feedback questionnaire following the day of evaluation, and evaluators' feedback questionnaires following the entire evaluation session.

a. *Four MESILA scores were studied*: scores on suitability to teacher-education studies, the teaching exercise, the interpersonal simulation, and the personal interview. Scores on the latter three components were rated individually for each task and are independent of each other. The score on suitability to teacher-education studies is a composite score of all the varied sub-scores rated throughout the test day. Scores range from a 1 to 9 scale. Mean scores and standard deviations for both populations are presented in Tables 1 and 2.

b. *Candidates' feedback questionnaire*: following the evaluation day, candidates were sent a feedback questionnaire, to be filled in anonymously. The questionnaires were distributed and collected from the FTF candidates before they left the evaluation facility, therefore almost all candidates (860) completed them. The remote candidates received the questionnaire by email, a week or two after the evaluation. Just over one-quarter of the candidates in the remote group returned the questionnaire (258). Candidates were asked about the evaluation day (was it respectful, organized, fair, and relevant to the world of teaching) and about the ability for each of the evaluation tools to diagnose the candidate. Answers were graded on a 1–5 scale.
c. *Evaluators' feedback questionnaire*: the evaluation team was asked to respond to a feedback questionnaire comparing remote evaluation to FTF evaluation. Two issues were addressed: the degree to which information was received in each type of evaluation, and the degree to which the candidates could be assessed in each.

7 Results

7.1 Candidates' evaluation scores on remote vs. face-to-face testing

The first question examined was the stability of the evaluation scores for the teacher-education programs, comparing the remote scores obtained in 2020 to the FTF scores of 2019. Means and standard deviations for the suitability to teaching scores by type of evaluation are presented in Table 1.

As seen in Table 1, a small, yet significant difference was found between mean suitability-for-teaching scores on the two evaluation methods, with scores being somewhat higher for the FTF tests ($M=6.89$, $SD=0.91$) than for the remote ones ($M=6.78$, $SD=0.90$; $t(1766)=2.59$, $p=0.01$). Cohen's $d=0.12$.

Table 1 Suitability for teaching scores from FTF and remote tests (*M* and SD)

Assessment method	Mean	SD	Skewness	Kurtosis	Number of candidates	Significance
FTF ($N=860$)	6.89	0.91	− 0.89	0.868	860	$t\,(1766)=-2.59$
Remote ($N=902$)	6.78	0.90	− 0.73	0.395	908	$p=0.01$, $d=0.12$

Table 2 Comparison of MESILA sub-scores: FTF and remote

Assessment method	Teaching exercise		Interpersonal simulation		Personal interview	
	M	SD	*M*	SD	*M*	SD
Face-to-face ($N=860$)	6.55	1.49	6.7	1.59	6.9	1.6
Remote ($N=902$)	6.22	1.52	6.4	1.56	6.7	1.4
Significance	$t\,(1759)=4.64$, $p=.00$		$t\,(1759)=3.98$, $p=.00$		$t\,(1759)=1.92$, $p=.05$	
Cohen's *d*	0.22		0.19		0.13	

Skewness scores indicate an upward trend for the high scores, and this trend is stronger in the FTF evaluation (− 0.89) than in the remote evaluation (− 0.73). Kurtosis scores are also indicative of more outlier scores in the FTF evaluation (0.868) than the remote evaluation (0.395).

Means and standard deviations for candidates' sub-test scores, for both evaluation methods, are presented in Table 2.

As seen in Table 2, a significant difference between both evaluation methods was found in all three MESILA components. Once again, the FTF scores were higher than the remote ones. This difference is greater for the teaching and simulation exercise ($d=0.22$ and 0.19), and smaller for the personal interview ($d=0.13$).

The internal structure of the MESILA exercises was examined to learn about the inter-relationships between the tools in both evaluation methods. Pearson correlations for these relationships are presented in Table 3.

As seen in Table 3, the relationships between the sub-tests differ for remote and FTF evaluation methods. In the FTF evaluation, there is a moderate and reasonable relationship (0.41–0.45) between the personal interview and the teaching and interpersonal simulation exercises. However, in the remote evaluation this relationship is stronger (0.73–0.79). An increase was also seen in the relationship between the teaching exercise and the interpersonal simulation in the remote evaluation, although to a lesser degree (increase from 0.81 to 0.89).

Table 3 Examination of the internal structure of MESILA exercises for FTF and remote assessment, using Pearson Correlations

Assessment method	Teaching exercise	Interpersonal simulation	Personal interview
FTF			
Teaching exercise	1		
Interpersonal simulation	0.81	1	
Personal interview	0.41	0.45	1
Remote			
Teaching exercise	1		
Interpersonal simulation	0.89	1	
Personal interview	0.73	0.79	1

All correlations are significant at $p=0.000$

The sharp increase in the correlations between the exercises may be due to the fact that in the FTF evaluation one pair of evaluators conducted the teaching and simulation exercises, and another conducted the personal interview, so that the scores were independent of each other. For logistic reasons, this format was changed in the remote evaluation, and the same pair of evaluators conducted both the exercises and the personal interview. This change seems to have reduced the independence of the exercises resulting in higher correlations between the scores.

To conclude this quantitative study, we have seen that the FTF evaluation scores are slightly higher than the remote scores, and the distribution of scores is wider. The remote evaluators tend not to use the extreme values on the scale, and center more on the middle range than did the FTF evaluators. Additionally, a significant increase is seen in the inter-correlations between the personal interview scores and the teaching and simulation exercises, which is indicative of a halo effect for the exercises, influencing the reliability of the scores obtained.

Next we will present the candidates' and evaluators' subjective perception of remote evaluation, compared to FTF evaluation.

7.2 Subjective perception of the quality of remote evaluation

7.2.1 Candidates' feedback questionnaire

After the evaluation day, candidates of both the FTF and the remote evaluation methods completed anonymous feedback questionnaires. The percentage of candidates who gave high scores (4 and 5) to statements about the evaluation is listed in Table 4. The questionnaires addressed two aspects of the evaluation, with the first three questions relating to the administration and content of the evaluation day, the next four questions examining the degree to which the evaluation allowed candidates to express themselves.

A chi-square test of independence was performed to examine the relation between method of evaluation and the candidates' subjective perception. As seen in Table 4, no significant differences were found between the FTF and remote evaluation regarding the organization of the test day or the relevance to the world of teaching. The findings are different regarding their ability to express themselves as measured by both evaluation measures. There was a significant decrease in their ratings of the fairness of the evaluation: 93.5% found the FTF evaluation to be fair, as compared to 80.2% for the remote evaluation. The candidates felt, significantly, that they are not as well expressed in two out of three of the evaluation components (teaching a subject matter and the simulation), although there was no difference in their evaluation of the interview by both methods. This finding testifies to a decrease in the faith candidates have in the remote evaluation system as opposed to FTF testing.

It is possible that these difference are a result of the different methods of gathering feedback—those who took the FTF tests filled in a paper questionnaire, which the evaluators handed to them at the end of the day. Although the questionnaire was anonymous, the candidates filled it in under the supervision of the evaluators, and it is possible that the candidates felt that their responses could affect their score. Conversely, those who underwent remote evaluation, received the questionnaire by email several days after the evaluation, and it was returned anonymously. Some of the candidates had received their scores before the questionnaires arrived, and were confident that the feedback could no longer affect these scores. It is possible that these candidates felt freer with their

Table 4 Candidates' subjective assessment of FTF and remote assessment (percent of high positive scores)

	Statement	FTF assessment ($N=860$)	Remote assessment ($N=258$)	Chi square
Administration of the assessment day	The assessment was respectful	98.9	95.3	X^2 (1, $N=1118$) = 14.79, $p<.001$
	The assessment was organized	89.8	90.3	X^2 (1, $N=1118$) = .01, $p=.949$
	The exercises are relevant to the world of teaching	83.3	84.5	X^2 (1, $N=1118$) = .65, $p=.4175$
The assessment allowed candidates to express themselves	The assessment was fair	93.5	80.2	X^2 (1, $N=1118$) = 31.25, $p<.001$
	I expressed myself in the interview	72.4	66.7	X^2 (1, $N=1118$) = 1.99, $p=.1574$
	I expressed myself in teaching subject matter	63.6	53.4	X^2 (1, $N=1118$) = 7.45, $p=.0063$
	I expressed myself in the simulation	76.9	63.5	X^2 (1, $N=1118$) = 15.72, $p<.001$

responses. Additionally, as not many candidates of the remote evaluation population returned the questionnaire, they may represent a self-selected, biased group.

7.2.2 Evaluators' feedback questionnaire

Because most evaluators had been part of the 2019 FTF evaluation and the 2020 remote evaluation, they were asked to respond to a feedback questionnaire comparing the two methods. The percentage of evaluators who gave high scores (4 and 5) to statements on both methods is listed in Table 5. The questionnaire addressed the evaluators' ability to obtain information and assess the candidates as well as their opinion on the convenience of the methods.

Table 5 Evaluators' objective perception of remote assessment ($N=14$)

Statement	%
Face-to-face assessment made for a better evaluation of the candidates	42.8
Remote assessment made for a better evaluation of the candidates	35.7
Remote assessment gives additional relevant information that is not obtained in FTF assessment	35.7
Lacking group exercise and group dynamics, critical information for evaluating the candidates is missing	35.7
It's possible to conduct a group assessment exercise over ZOOM	64.3
Remote assessment was more convenient than coming to the site and conducting FTF assessment	85.7

The data reveals a lack of consensus among the evaluators who participated in both types of evaluation regarding the preferred method to professionally assess the candidates' attributes for teacher-education studies.

a. Quality of evaluation: about 43% prefer FTF evaluation, but 36% preferred remote evaluation and felt that it provides some relevant information that is not obtained in FTF evaluation.
b. Absence of the group dynamics exercise in the remote evaluation: about one third of the evaluators did not miss this exercise. Two thirds of the evaluators think that a group exercise can be conducted in remote evaluation as well.
c. Convenience: the vast majority found remote evaluation to be more convenient.

To conclude the subjective perception of the participants, a certain decrease can be seen in the candidates' perception of the diagnostic abilities of the evaluation and its fairness when compared to the FTF evaluation. There is no consensus among evaluators as to differences between the two methods, although most of them feel that remote evaluation is more convenient and that various tools, such as group tests, can be added.

8 Discussion

The present study was designed to examine the effect of the transition from FTF evaluation to remote evaluation on the quality of candidates for teacher-education studies. To do so, we analyzed the information we had accumulated from the FTF evaluation

conducted in 2019 and the remote evaluation conducted in 2020. Three levels of the evaluation were analyzed: quality of scores, candidates' subjective perception of each method and the evaluators' perception regarding the comparison between the two methods.

Our examination of the quality of the remote evaluation scores revealed a similarity to score attributes found for FTF testing. Despite the slight decline in mean scores in the remote evaluation, the score distribution resembles that of the previous year, when testing was face-to-face, with a slight increase in centralization of the scores. An examination of the internal structure of the exercise scores reveals a higher correlation between the tools in the remote evaluation, perhaps because the same evaluation team worked throughout all exercises, and not having an independent team for the interview, as in the FTF evaluation.

Examining the participants' subjective perception of the remote evaluation we found that the candidates gave the remote teacher-education evaluation high scores for the administration of the evaluation day, however, there was a certain decrease in the candidates' faith in their ability to express themselves during the remote evaluation. It was postulated that this finding might be due to difference in the methods of gathering the feedback questionnaires.

Finally, we asked the evaluators (most of whom had had the experience of both FTF and remote evaluation) to compare the two evaluation methods. We found that there was no agreement as to the preferred method—some preferred the first, others preferred the second. The team responded that even though the tests are conducted remotely, in their opinion the interactive group tests could be reinstated. One clear and unequivocal finding was that remote evaluation is preferred for its convenience.

The findings obtained in the present study are consistent with those cited in the literature (Basch et al., 2021; Blacksmith et al., 2016; Sears et al., 2013) in which interviewees received lower performance ratings in videoconference interviews than in face-to-face (FTF) interviews and interviewees held more negative perceptions of these interviews.

Several reasons can be suggested for the lower scores on the remote test. One observed phenomenon is that the distribution of the scores tends to be central on the remote test. Perhaps the evaluators feel less confident about the information obtained, they have fewer indicators that enable them to give very high or very low grades, and their scores therefore concentrate mid-scale. The absence of interpersonal group exercises, as well as the absence of peer evaluations decreases the amount of information that the evaluators have and the number of opportunities to observe unusual—both negative and positive—behaviors. It has been found that evaluators' lack of confidence decreases as they gain experience with remote evaluation (Helding, 2020). In the future, re-introducing interactive group tests and peer evaluations in the evaluation system will provide additional information about the candidates and may enhance the evaluator's confidence in their ratings.

A second possible factor relates to emotional, social and health differences between the two populations. The 2020 population were in the throes of the COVID-19 pandemic with accompanying mental stress and physical influences which could have significantly affected the assessment scores.

Another possible reason for the lower scores for remote evaluation is related to the effect of this evaluation method on the candidates. As seen in the Literature survey, scores on remote tests can be affected by technological factors (Denter & George, 2020), psychological attributes of the diagnostic situation (Helding, 2020; Weinberg &

Rolnick, 2020), or social factors such as the lack of a comparison group of other candidates during the evaluation (Schlapobersky, 1993). These effects could have led to lower performance on the remote evaluation than on the FTF one, reflected in somewhat lower scores.

The catalyst for the introduction of remote testing was the COVID-19 pandemic and the required social distancing. However, we can take advantage of this period to learn of the affordances and constraints of remote testing as opposed to FTF testing.

1. Remote testing as a reflection of teaching skills: digital pedagogy and the incorporation of on line teaching are of growing importance for teachers in the twenty-first century (Rudolph et al., 2022). Evaluating candidates using remote computerized testing can be used to help evaluate the candidates' ability to adapt to this media and their potential to teach effectively online.
2. Concentration on the essence: in remote testing by Zoom technology the evaluator concentrates mainly on the candidate's face and voice. Body language, hand and leg movements and other subconscious clues are not usually seen, reducing their contribution to the personality evaluation. However, one benefit of this is the concentration on the essence of the candidate's answers, without inconsequential distractions. The evaluators are required to listen more carefully to what is being said and to pay more attention to the candidate's facial expressions without extraneous distractors (Weinberg & Rolnick, 2020).
3. Bias reduction: interpersonal exercises and interviews are often influenced by tester bias resulting from unrelated factors such as the candidate's origins, looks, dress, posture, etc. Remote testing neutralizes much of this bias by concentrating on the candidate's answers. This should improve the reliability of the scores, as well as increasing fairness and equality (Woods et al., 2020).
4. Recording the session: remote testing enables recording the testing session with the candidate's permission. This could add many benefits which are lacking in FTF testing. Such as: (a) feedback to the evaluators—the supervisor can study a sample of exercises and interviews to give feedback and improve the evaluators' technique. (b) Inter-rater reliability can be measured and improved by having a group of evaluators watch the recording and rate the candidates simultaneously, (c) scoring and appeals—questions about scoring, appeals or borderline cases can be re-evaluated by the evaluator or his superior based on the recording, (d) recorded tests can be used for training new evaluators or during advanced training, (e) recorded sessions can be analyzed for research purposes to improve the quality of the exercises, interview questions or the evaluation process.
5. Remote technology has new and updated features which can be incorporated into the testing situation. For example, Zoom allows participants to use a whiteboard or shared screen, or to enter break-out rooms, which can be used for dyadic exercises, or for reshuffling the group assignments during an exercise to test such dimensions as flexibility and adapting to change. Such instructions during FTF testing are often difficult to implement logistically.
6. A disadvantage of remote testing is the requirement of access to computer, camera, microphone and internet connections which increases inequity between groups of different socioeconomic levels (Vadtal & Georgi, 2020). We solved this by providing access to those who did not have these. On the other hand, remote testing provides equal opportunity and accessibility for candidates living far away, disabled candidates or those

with special needs who are discriminated against by the need to travel to FTF tests in limited test sites.

Thus we see that beyond the questions studied in this paper, there are many facets of remote versus FTF evaluations that may influence decisions as to the method employed to evaluate candidates.

9 Conclusions

As Basch et al. (2021) warned in the discussion of practical implications to their study of FTF and videoconferences, "organizations should not overlook that the interview medium affects applicants' chances in a selection process" (p. 935).

The above discussion related to some benefits and disadvantages of remote evaluations for the candidates and for the evaluators. Beyond the issue of the feasibility of using remote MESILA tests, there is the question of their predictive validity, namely, the degree to which the candidates' performance on the remote tests reflect their suitability to teacher-education studies. Significant predictive validity has been found for MESILA FTF tests with criteria of success in academic teacher-education studies, pre-service training ratings, classroom teaching evaluations and receiving tenure (Goldenberg, 2020). The similarity in score attributes for both methods suggests initial indications that the remote evaluation will preserve the predictive validity found in the FTF evaluation. However, only a future study of validity could indicate whether an evaluation system measured by remote means could predict success in teacher-education studies and integration into the field of teaching.

10 Research limitations

The present study is an examination of the remote scores of all 2020 candidates for teacher-education studies. At the beginning of the evaluation season the team was introduced to a new evaluation method, with unfamiliar technology and new operating instructions. Up to then, the evaluators' role focused only on diagnosing the candidates in the group. Their role was now expanded to include explaining the use of Zoom to the candidates during the exercises, simulations, and interviews. They also had to deal with such issues as computer problems that came up during the evaluation and instructing the candidates how to leave and enter the system before and after breaks. These roles were all new to members of the evaluation team who had to learn them as they were adjusting to the new reality. In addition, group interactions and peer assessments were removed from the remote test battery, adding additional changes for the evaluators. These limitations may explain part of the significant differences noted in this research. Perhaps the study should have included just a sample of candidates who were tested after the team had a few months to adjust to the remote method. It is recommended that the scores for the 2021 evaluation be examined, after the evaluators had time to integrate and familiarize themselves with the remote system.

Declarations

Ethical consent All candidates undergoing the MESILA testing give their written consent for the use of their test scores for research purposes.

References

Basch, J. M., Melchers, K. G., Kegelmann, J., & Lieb, L. (2020). Smile for the camera! The role of social presence and impression management in perceptions of technology-mediated interviews. *Journal of Managerial Psychology, 35*, 285–299. https://doi.org/10.1108/JMP-09-2018-0398

Basch, J. M., Melchers, K. G., Kurz, A., Kreiger, M., & Miller, L. (2021). It takes more than a good camera: Which factors contribute to differences between face-to-face interviews and videoconference interviews regarding performance ratings and interviewee Perceptions? *Journal of Business Psychology, 36*, 921–940. https://doi.org/10.1007/s10869-020-09714-3

Blacksmith, N., Wilford, J. C., & Behrend, T. S. (2016). Technology in the employment interview: A meta-analysis and future research agenda. *Personnel Assessment and Decisions, 2*, 12–20. https://doi.org/10.25035/pad.2016.002

Casey, C., & Childs, R. (2011). Teacher education admission criteria as measure of preparedness for teaching. *Canadian Journal of Education, 34*(2), 3–20.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*, 554–571. https://doi.org/10.1287/mnsc.32.5.554

Darling-Hammond, L., & Hyler, M. E. (2020). Preparing educators for the time of COVID … and beyond. *European Journal of Teacher Education, 43*(4), 457–465. https://doi.org/10.1080/02619768.2020.1816961

Ekman, P. (2004). Emotional and conversational nonverbal Signals. In J. M. Larrazabal & L. A. P. Miranda (Eds.), *Language, knowledge, and representation. Philosophical Studies Series.* (Vol. 99). Springer. https://doi.org/10.1007/978-1-4020-2783-3_3

Fincham, J. F. (2008). Response rates and responsiveness for surveys, standards, and the journal. *American Journal of Pharmaceutical Education, 72*(2), 43. https://doi.org/10.5688/aj72024315

Fullwood, C., & Finn, M. (2010). Video-mediated communication and impression formation: An integrative review. In A. C. Rayler (Ed.), *Videoconferencing: Technology, impact, and applications* (pp. 35–55). Nova Science Publishers.

Goldenberg, J. (2018). MESILA—A selection battery for future teachers. *International Forum for Teacher Development.* Retrieved from https://info-ted.eu/mesila-a-selection-battery-for-future-teachers/

Goldenberg, J. (2020). *MESILA screening tests for teacher-education studies candidates.* MOFET Institute (Hebrew).

Helding, L. (2020). Cognition in the age of corona: Teaching students how to learn. *Journal of Singing, 77*(2), 249–259.

Holmes, C. M., & Kozlowski, K. A. (2015). A preliminary comparison of online and face-to-face process groups. *Journal of Technology in Human Services, 33*(3), 241.

Horn, R. G., & Behrend, T. S. (2017). Video killed the interview star: Does picture-in-picture affect interview performance? *Personnel Assessment and Decisions, 3*, 51–59. https://doi.org/10.25035/pad.2017.005

Klassen, R. M., & Kim, L. E. (2019). Selecting teachers and prospective teachers: A meta-analysis. *Educational Research Review, 26*, 32–51. https://doi.org/10.1016/j.edurev.2018.12.003

Manella, M. (2020). From video interviews, through cancelling meetings, to freezing recruitment: Work in Israel in the shadow of COVID-19. *Calcalist* (Hebrew). https://www.calcalist.co.il/local/articles/0,7340,L-3799476,00.html

McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., & Ahmed, S. M. (2017). Applicant perspectives during selection: A review addressing "So what?", "What's new?", and "Where to next?" *Journal of Management, 43*, 1693–1725.

Mihelic, G., Bosch, C., Boyd, K., & Miller, M. L. (2018). The relationship between admission criteria and preservice teacher preparedness for a small rural educator preparation provider, *Administrative Issues Journal, 8*(2). https://dc.swosu.edu/aij/vol8/iss2/5

Potosky, D. (2008). A conceptual framework for the role of the administration medium in the personnel assessment process. *Academy of Management Review, 33*(3), 629–648.

Potosky, D., & Bobko, P. (2004). Selection testing via the internet: Practical considerations and exploratory empirical findings. *Personnel Psychology, 57*, 1003–1034. https://doi.org/10.1111/j.1744-6570.2004.00013.x

Roberts, F., & Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *Journal of the Acoustic Society of America, 133*(6), L471–L477.

Roloff, J., Klusmann, U., & Lüdtke, O. (2020). The predictive validity of teachers' personality, cognitive and academic abilities at the end of high school on instructional quality in Germany: A longitudinal study. *AERA Open, 6*(1), 1–17. https://doi.org/10.1177/2332858419897884

Rudolph, J., Tan, S., Crawford, J., et al. (2022). Perceived quality of online learning during COVID-19 in higher education in Singapore: Perspectives from students, lecturers, and academic leaders. *Educ Res Policy Prac*. https://doi.org/10.1007/s10671-022-09325-0

Russell, G. I. (2015). *Screen relations: The limits of computer-mediated psychoanalysis and psychotherapy.* Karnac.

Schlapobersky, J. (1993). The language of the group: Monologue, dialogue, and discourse in group analysis. In D. Brown & L. Zinkin (Eds.), *The psyche and the social world: Developments in group-analytic theory* (pp. 211–231). Routledge.

Schmidt, F. L. (2016). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 Years.* Working Paper October 2016. https://doi.org/10.13140/RG.2.2.18843.26400

Sears, G., Zhang, H., Wiesner, W., Hackett, R., & Yuan, Y. (2013). A comparative assessment of videoconferencing and face-to-face employment interviews. *Management Decision, 51*, 1733–1752. https://doi.org/10.1108/MD-09-2012-0642

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunication*. Wiley.

Sklar, J. (2020). 'Zoom fatigue' is taxing the brain. Here's why that happens. *National Geographic24*.

Slovák, P. (2007). Effect of videoconferencing environments on perception of communication. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *1*(1).

Straus, S. G., Miles, J. A., & Levesque, L. L. (2001). The effects of videoconference, telephone, and face-to-face media on interviewer and applicant judgments in employment interviews. *Journal of Management, 27*, 363–381.

Vadtal, L., & Georgi, A. (2020). We are a very embarrassing startup nation—and that's an outrage (Hebrew). https://www.themarker.com/markerweek/MAGAZINE-1.8794241

Wegge, J. (2006). Communication via videoconference: Emotional and cognitive consequences of affective personality dispositions, seeing one's own picture, and disturbing events. *Human-Computer Interaction, 21*(3), 273–318.

Weinberg, H., & Rolnick, A. (2020). *Theory and practice of online therapy: Internet-delivered interventions for individuals, families, groups, and organizations*. Routledge.

Wolf, C. R. (2020). Virtual platforms are helpful tools but can add to our stress. *Psychology Today*. May 14. Retrieved October 19, 2020, from https://www.psychologytoday.com/us/blog/the-desk-the-mental-health-lawyer/202005/virtual-platforms-are-helpful-tools-can-add-our-stress

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology, 29*, 64–77.

Yun, G. W., & Trumbo, C. W. (2000). Comparative response to a survey executed by post, e-mail, and web form. J Compu-Mediated Com. Retrieved April 7, 2008, from http://jcmc.indiana.edu/vol6/issue1/yun.html