Assessment *for* Graduate Teaching Report 2021

Prepared for the Assessment *for* Graduate Teaching (A*f*GT) Consortium Collaborators & Licensees

October 2021

Title Page

Table of Contents

Executive Summary

Introduction

Consortium Update

Findings from 2020 Data

Moderation and Evaluation

Instrument Refinement

Consortium Initiatives

References

Acknowledgements

The Centre for Program Evaluation (CPE), as the Management Group of the A*f*GT Consortium, would like to acknowledge the substantial work of all members of the A*f*GT Consortium who have enthusiastically participated in the design, piloting, trialling and implementation of the A*f*GT, and who have also actively participated in evaluation and moderation activities upon which this report relies.

We also wish to acknowledge the input provided by personnel from the Centre for Program Evaluation:

Mia Chen
Megan Dennis
Naoki Ikeda
Megan Stonnill

Recommended citation:

Keamy, K., Clinton, J., & Tan, K., (2021). *Assessment for Graduate Teaching Report 2021: Prepared for AfGT Consortium Collaborators*. A*f*GT Consortium, Melbourne Graduate School of Education, The University of Melbourne.

Contact Details:

Email: afgt-help@unimelb.edu.au
Centre for Program Evaluation, Melbourne Graduate School of Education
The University of Melbourne
Parkville 3010 VIC

Image by Taylor Wilcox from Unsplash

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

Table of Contents

Executive Summary 5
1. Introduction 11
2. Consortium Update 12
2.1. Consortium Members 12
2.2. Response to COVID-19 12
2.3. Jurisdictional and regulatory responses 17
2.4. AfGT Sustainability Plan 18
2.5. Governance Arrangements 18
2.6. Activities of the Consortium and Executive Group 20
2.7. Benchmarking Project (AITSL) 22
2.8. Publications/Conferences 22
3. Findings From 2020 Data 24
3.1. About the AfGT instrument 24
3.2. Findings 25
3.3. Scale Reliability and Instrument Validation 28
3.4. Validation of the Cut Score 39
4. Moderation and Evaluation 40
4.1. Moderation of AfGT 40
4.2. AfGT Process Evaluation 45
4.3 Summary and Future Considerations for the Instrument 54
5. Refinements to the Instrument 57
6. Consortium Initiatives 58
6.1. The Use of Computers to Support Assessors 58
6.2. Resources to Support Schools and Mentor Teachers 58
6.3. Providing Institutions with Assessment Feedback 58
6.4. Moving Ethics Documentation to Online 58
7. References 59

List of Tables

Table 1. AfGT COVID-19 Decision Making Matrix 16
Table 2. Summary of jurisdictional and regulatory responses to COVID-19 17
Table 3. Entities Within the AfGT and Meeting Frequency 20
Table 4. Committee achievements since last report 21
Table 5. Participant Demographics 25
Table 6. Mean scores and SD by element 25
Table 7. Mean Scores and SD by Each Element by Institutions 27
Table 8. Factor Solution with Estimate of Reliability 29
Table 9. Goodness-of-Fit Statistics 30
Table 10. Item Statistics 31
Table 11. Descriptive Summary by Element for Bachelor and Masters program 36
Table 12. Descriptive Summary by Element for program type 36
Table 13. DIF output for Primary vs Secondary program type 37
Table 14. Descriptive Summary of Cut Score by Element 39
Table 15. Moderation Workshops Details 40
Table 16. Descriptive Statistics of Moderation Data 42
Table 17. Feedback on AfGT's overall impact on teaching practice 46
Table 18. Feedback on AfGT's impact on PSTs' professional learning 46
Table 19. Survey Responses from PSTs 49
Table 20. Summary of Refinements to the AfGT Documentation 57

List of Figures

Figure 1. 2020 Timeline of COVID-19 disruptions to teaching, learning and research 13
Figure 2. AfGT COVID-19 Decision Making Tree 15
Figure 3. AfGT Consortium governance structure 19
Figure 4. AfGT assessment task summary 24
Figure 5. Grade distribution by element 26
Figure 6. Average scores by each element by institution 27
Figure 7. Correlation matrix of AfGT items 28
Figure 8. Factor analysis using three-factor and four-factor structures 29
Figure 9. Category Characteristics Curves for Element 1 33
Figure 10. Category Characteristics Curves for Element 2 33
Figure 11. Category Characteristics Curves for Element 3 33
Figure 12. Category Characteristics Curves for Element 4 33
Figure 13. Test Information Function for Element 1 to Element 4 35
Figure 14. Category Characteristic Curve for Element 1 Item 5 (Primary vs Secondary) 38
Figure 15. Internal consistency of moderation data - assessors' view 43
Figure 16. Internal consistency of moderation data - item view 44
Figure 17. Support received by teacher educators and placement officers 47
Figure 18. Distribution of placement days 49
Figure 19. Time taken to complete each element 49
Figure 20. Clarity of task 44
Figure 21. Relevance of task 50
Figure 22. Degree of difficulty 50
Figure 23. Usage of AfGT materials 51
Figure 24. Clarity of materials and coherence of assessment 51
Figure 25. Unforeseen events identified by PSTs that impacted their completion of the AfGT 52
Figure 26. Framework for establishing AfGT's assessment validity and reliability 52

List of Abbreviations

Abbreviation	Full text
AARE	Australian Association for Research in Education
ACECQA	Australian Children's Education & Care Quality Authority
AITSL	Australian Institute for Teaching and School Leadership
A*f*GT	Assessment *for* Graduate Teaching
ANOVA	Analysis of Variance
APST	Australian Professional Standards for Teachers
CCC	Category Characteristic Curve
CIM	Cross-Institutional Moderation
CRT	COVID-19 Response Team
DET	Department of Education and Training
DIF	Differential Item Functioning
DoE	Department of Education
EAG	Expert Advisory Group
FAQ	Frequently Asked Questions
GPCM	Generalised Partial Credit Model
GRM	Graded Response Model
GTPA	Graduate Teacher Performance Assessment
GTS	Graduate Teacher Standards
IRT	Item Response Theory
ITE	Initial Teacher Education
LMS	Learning Management System
PST	Pre-Service Teacher
QTPA	Quality Teaching Performance Assessment
RISEC	Research in Schools and Early Childhood Settings Victoria
TIF	Test Information Function
TPA	Teaching Performance Assessment
TRB	Teacher Registration Board
TRBWA	Teacher Registration Board of Western Australia
UoC	The University of Canberra
UoM	The University of Melbourne
VIT	Victorian Institute of Teaching

Executive Summary

Introduction

The A*f*GT Consortium was founded in 2017 as part of an Australian Institute for Teaching & School Leadership (AITSL) seed-funded project to develop a teaching performance assessment. The original project was completed in 2018 and the A*f*GT Consortium operates now as a self-governed and self-funded body. In May 2018, AITSL's Expert Advisory Group advised that the A*f*GT instrument designed and developed by the A*f*GT Consortium is a valid instrument for assessing whether a pre-service teacher's performance meets the *Australian Professional Standards for Teachers* (APSTs) at the Graduate Teacher level. The Expert Advisory Group re-iterated this advice in July 2019, following the provision of data based on a larger data set in the following year.

Consortium Update

The following institutions are Consortium Collaborators in the A*f*GT Consortium:

- University of Melbourne (Lead Institution)
- Charles Darwin University
- Curtin University
- Federation University
- University of Canberra
- University of Sydney
- University of Western Australia
- University of Technology Sydney
- Victoria University

The following institutions are Consortium Licensees:

- Montessori Institute, Western Australia (commenced early 2019)
- Excelsia College, Sydney (commenced early 2020)
- Melbourne Polytechnic (commenced mid 2021)
- Southern Cross Education Institute (commencing start 2022)
- University of Adelaide (commencing start 2022)

The A*f*GT governance structure remain unchanged in 2020/2021. However, in May 2020, the governance document was revised to ensure that the positions of Chair and Deputy Chair of the Executive and Consortium were fully described, along with the roles of the Director and Project Manager of the A*f*GT Management Team. The impetus for these revisions was in relation to succession planning so that there is a framework to guide future governance and management of the A*f*GT Consortium.

In response to the COVID-19 pandemic, the Executive Group of the A*f*GT Consortium established a COVID-19 Response Team (CRT) to assist institutions in their implementation of the A*f*GT during and following the COVID-19 crisis. Terms of Reference for the CRT were established, and a Decision-making Package was developed, including a Decision-making Tree, to assist institutions to make their own decisions in relation to the implementation of the A*f*GT, including adaptations they might need to make, whilst maintaining the integrity of the instrument. The package was developed collaboratively with input from members.

Image by Hermann Traub from Pixabay

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

In addition to the Consortium's Executive Group and the AfGT Management Group, which provides operational and administrative support, the Consortium is further supported by five Committees, namely:

- Assessment & Measurement Committee (AMC)
- Ethics & Privacy Committee (EPC)
- Implementation & Improvement Committee (IIC)
- Research & Publication Committee (RPC);
- Promotion & Induction Committee (PIC) ad hoc

Conference presentations and joint publications are the main means by which the Consortium shares insights gained in all aspects of the design, trial and implementation of the A*f*GT with the sector. The Research & Publications Committee has developed documentation to record the planned and completed conference presentations and publications and this has been used to guide a steady increase in a number of articles for publication. Due to COVID-19, many conferences were cancelled in 2020. However, Consortium members remained active, and as a result, three articles from collaborations of academics from more than half of the institutions in the Consortium were accepted and published by high-quality journals.

Findings from 2020 Data

As shown in Figure A on the right, the A*f*GT comprises four elements, each containing several inter-related tasks.

Overall, 2348 PSTs completed the A*f*GT across eleven institutions in 2020. Consistent with prior years, there were significantly more female PSTs (65%) in the 2020 cohort compared to males (30%) and other genders. The breakdown between undergraduate and postgraduate PSTs, which is determined by the programs offered by the respective institutions, was almost equal between masters (47%) and bachelor programs (53%). Within the bachelors, the largest cohort was the Bachelor Primary (24%), whereas the largest Masters cohort was the Masters Secondary (36%).

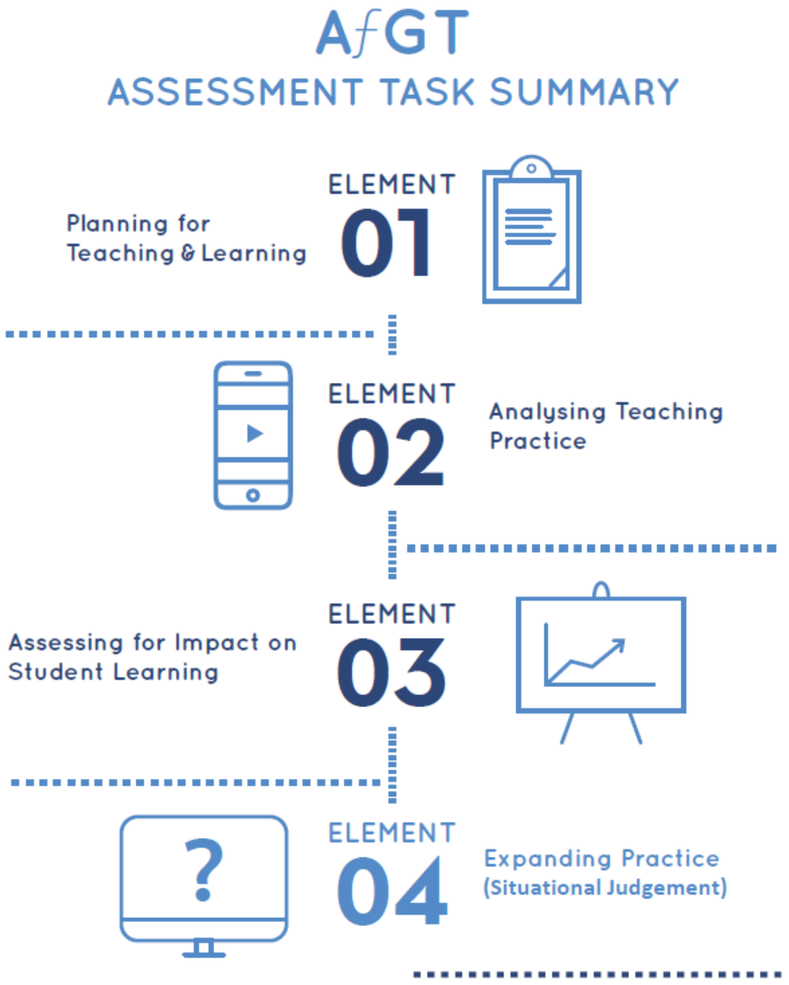


Figure A. Overview of the four elements that comprise the A*f*GT

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

To understand the average and distribution of participants' scores, the mean and standard deviation of scores for each element were calculated. The results demonstrate high consistency in the distribution of grades across the four elements, suggesting consistency in scoring across the elements. The data also suggest that there is an opportunity for success in each of the elements and a relatively equivalent score across the elements. This is significant, and provides support that the instrument is highly stable, given the varying number of tasks (or items) in each element. The distribution of grades for each element across each institution is fairly consistent, although there is some variability between each element for three of the institutions.

The data were further analysed for reliability, consequential process evaluation, validity and fairness as summarised in Table B.

	2019 *(%)*	2020 *(%)*
Gender		
Female	1255 *(75%)*	1528 *(65%)*
Male	421 *(25%)*	697 *(30%)*
Other	-	1 (0%)
Missing Data	-	122 (5%)
Program Type		
Bachelor Early Childhood	96 *(6%)*	107 *(5%)*
Bachelor Primary	373 *(22%)*	552 *(24%)*
Bachelor Secondary	268 *(16%)*	282 *(12%)*
Bachelor EC/Primary	48 *(3%)*	27 *(1%)*
Bachelor Primary/Secondary	-	267 *(11%)*
Masters Early Childhood	-	12 *(0%)*
Masters Primary	179 *(11%)*	207 *(9%)*
Masters Secondary	646 *(39%)*	857 *(36%)*
Masters EC/Primary	66 *(3%)*	37*(2%)*
TOTAL	**1676**	**2348**

Table A. Participant Demographics

Evidence Type	Analysis	Purpose of Analysis
Reliability Evidence	Inter-rater reliability	Determine consistency of judgement among assessors using moderation data
	Cronbach's alpha	Measure of internal consistency
Process Evaluation (Consequential)	Descriptive statistics and Qualitative analysis	Feedback from participants collected via survey and interview/focus group data
Validity Evidence	Descriptive statistics	Determine distribution, central tendency and dispersion of data
	Factor analysis	Determine if tasks making up the four elements group together as theorised, indicating each element as independent factors that measure unique aspects of teacher readiness
	Correlations	Evaluate the strength of relationships between tasks and elements
	Item Response Theory (IRT) analysis	Evaluate how well an assessment and items within an assessment work
	Test information	Indicates how well an assessment estimates a PST's location on a performance scale
Fairness Evidence	Descriptive statistics	Provide statistical data by gender, program type and program specialisation
	t-test	Determine if two groups are statistically different from each other
	ANOVA	Determine if three or more groups are statistically different from each other
	Differential Item Functioning (DIF)	Identify presence of potential bias in as assessment with respect to a PST belonging to a specific group (Bachelor vs Masters program type or Primary vs Secondary program specialisation)

Table B. Summary of Data Analysis

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

Cross-institution Moderation

Consistent with previous years in which a mixed method approach with ongoing validation was adopted, the Consortium continued its cross-institution moderation exercises and collected evaluation data in 2020. Due to the COVID-19 pandemic, the cross-institution moderation exercises were conducted fully online and took place in November 2020 and February 2021. It was necessary to adjust the timing due to the disruptions faced in placements of PSTs as schools were either closed or had switched to remote or dual learning modalities. As a consequence, a significant number of PSTs' placements were delayed and pushed towards the last quarter of 2020 resulting in the A*f*GT assessment data not being finalised until the end of 2020 and the first quarter of 2021.

Cross-institution moderation activities are one of two moderation dimensions of the A*f*GT. Prior to the cross-institutional exercise, each institution conducted their internal moderation activities within and across their program of study in accordance with their university policies to ensure the continuous fidelity and validation of the A*f*GT instrument. The online cross-institution moderation workshops are designed to collaboratively engage the whole Consortium in the moderation process, while at the same time determining any revisions that might need to be made to the A*f*GT.

Two main observations were made following the moderation exercises:

- there is significant improvement in the degree of agreement among the assessors as the moderation rounds progressed from November 2020 to February 2021, and
- based on inter-rater reliability measures, there is strong evidence to suggest that assessors agree what classroom readiness looks like and the performance standard that meets the APST at Graduate level.

Process Evaluation

The 2020 process evaluation utilised a mixed methods approach and evaluation data was collected through online surveys and one focus group. Three main participant groups of teacher educators, placement officers and PSTs were invited to participate in the evaluation process. Due to the small number of participants who responded from the placement officers' group, their evaluation data are combined with the teacher educator group when the findings are reported. The process evaluation is designed to collect information on the implementation of the instrument and any associated challenges faced by stakeholders directly involved in the implementation of the A*f*GT.

Teacher Educators and Placement Officers

Overall, despite the challenges faced in 2020 due to the COVID-19 pandemic, teacher educators and placement officers regarded the A*f*GT as a valid and coherent teaching performance assessment instrument. It is notable that the feedback is more precise around specific implementation challenges and areas within the instrument which require attention.

The level of insight and familiarity with the requirements of the A*f*GT developed by assessors provides support that the A*f*GT is an established, mature assessment that is subject to continuous review and evaluation. Opportunities for more advanced resources such as annotated examples and capacity building infrastructure, such as assessor training may be considered by the Consortium in the future.

Pre-service Teachers

The survey for PSTs explored more detailed aspects of the A*f*GT from a user perspective including the clarity, relevance and difficulty of each A*f*GT element as well as PSTs' feedback on the guidance materials provided. A total of 87 PSTs responded to the survey in 2020, representing both undergraduate (46 PSTs) and postgraduate (41 PSTs) degrees.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

For all the dimensions surveyed, the 2020 PST data were more favourable compared to previous years. In 2020, more PSTs agreed that the A*f*GT was clear, relevant, coherent and had a more manageable degree of difficulty in the tasks. When analysed together, PSTs found the A*f*GT tasks to be both coherent and challenging. They also perceived the A*f*GT assessment as relevant and an appropriate indicator of their classroom readiness.

When asked to identify any unforeseen events that may have hindered or interrupted PSTs' completion of the A*f*GT, an overwhelming majority identified COVID-19 as a major challenge. Whilst the A*f*GT continued to be implemented with fidelity throughout the challenging COVID-19 pandemic in 2020, some PSTs had to adapt to a shorter placement period. This had a negative impact on PSTs' professional experience and many expressed the shorter placement was an added challenge to completing the A*f*GT.

Photo by Jonathan Cosens Photography on Unsplash

Summary and Future Considerations for the Instrument

The findings arising from these analyses are arranged here in relation to the elements for verification in Program Standard 1.2 (AITSL, 2019):

1. **Valid reflection of classroom teaching practice (including planning, teaching, reflecting and assessing student learning):**
   a. The results reveal that the A*f*GT is a valid reflection of classroom teaching and that the majority of the A*f*GT items are correctly ordered.
   b. Collectively, the results demonstrate high consistency in the distribution of grades across the four elements, suggesting consistency in scoring across the elements.
   c. The data suggest that there is an opportunity for success in each of the elements and a relatively equivalent score across the elements. This is significant and provides support that the instrument is robust and highly stable, given the varying number of tasks in each element.
   d. The A*f*GT is not showing any systematic bias for the various sub-groups of program type (bachelor, masters, primary, secondary, early childhood, etc.), although the data reveals that PSTs from primary and secondary program type score slightly differently to each other on a small number of tasks in Elements 1 and 3.
   e. When teacher educators were asked if the A*f*GT adequately measures the planning, teaching, assessing and reflecting aspects of teaching practice, 75% of the respondents said yes.
2. **Valid assessment that assesses the content of the Graduate Teacher Standards:**
   a. Given the objective of the A*f*GT is to assess PSTs' attainment of the specified APSTs at the Graduate level rather than used as a ranked assessment, the results reveal that the conceptual design of the A*f*GT is a valid assessment of the content of the Graduate Teacher Standards.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

b. To pass the A*f*GT, PSTs are required to pass all four elements. While the A*f*GT assesses the content of all the Graduate Teacher Standards, it is possible to identify the items that prove to be most and least challenging to achieve a 'G' or 'G+'.

c. When PSTs were asked how relevant the A*f*GT tasks were in reflecting the Graduate Teacher Standards, 71% responded favourably.

3. Measurable and justifiable achievement criteria that discriminate between meeting and not meeting the Graduate Teacher Standards:

a. For all four elements, the A*f*GT is highly effective at obtaining precise estimates of PSTs who are 'on-the-cusp', where it is critical to determine if a PST has indeed met the APSTs at Graduate level. Element 4 reflects this particularly well.

b. The A*f*GT items are effective in separating PSTs at the low end of the classroom readiness scale.

c. As part of the ongoing validation process, the cut score was confirmed as representative of the score distribution based on 2020 sample data.

4. Reliability of scoring between assessors:

a. The distribution of grades for each element across each institution is consistent, with some within-institution variations identifiable.

b. Overall, all the assessors who participated in the cross-institution moderation process showed high internal consistency in their marking.

c. There is better strength of agreement for higher performing scripts relative to low performing scripts, with more variability for low performance submissions.

5. Moderation processes that support consistent decision making against achievement criteria:

a. Consistent with prior years, the inter-rater reliability analysis showed strong consensus among the assessors who participated in the standard-setting activity. Importantly, the assessors achieved stronger levels of agreement as the moderation rounds progressed through the online cross-institution moderation workshops.

b. There is strong evidence to suggest that assessors agree what classroom readiness looks like and the performance standard that meets the APST at Graduate level.

c. The current cross-institution moderation process ensures high-quality data are gathered as evidence of validity and reliability of the instrument whilst informing the Consortium on specific areas in the instrument which may benefit from refinement. This ensures the task descriptions remains clear and are contextually responsive as part of the continuous improvement process.

d. Maintaining vigilance on the cross-institution moderation processes will remain a high priority for the Consortium.

The results and processes contributed to the validation of the A*f*GT instrument. Based on 2020 data, the analyses continue to substantiate the A*f*GT as a valid, reliable and fair teaching performance assessment instrument. At an instrument-level, the A*f*GT was robust and coherent, and at item-level, the A*f*GT demonstrated well-ordered statistic parameters with strong and reliable test information.

These are characteristics of a mature, large-scale assessment with established merit and utility. Given that this data were collected in 2020 when the assessment was undertaken amidst the COVID-19 pandemic, speaks to the robustness and agility of the instrument. However, the instrument is not implemented in a vacuum. Maintaining the fidelity of the A*f*GT is only possible through the resolve and hard work of Consortium institutions and its people. It is not an insignificant achievement on the part of the Consortium to have achieved so much in such challenging and disruptive contexts.

Moving forward, a number of areas have been identified as key focus areas for the A*f*GT Consortium in the next three to five years. These include several initiatives that were put on hold due to the pandemic, which are anticipated to resume in the coming months, namely developing resources to support Consortium members, providing individual institutions with customised data analysis specific to their institution, and moving ethics documentation to an online format.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

1. Introduction

The A*f*GT Consortium was founded in 2017 as part of an AITSL-seed funded project to develop a teaching performance assessment project. The original project was completed in 2018 and the A*f*GT Consortium operates now as a self-governed and self-funded body.

In May 2018, the Expert Advisory Group advised that the A*f*GT instrument designed and developed by the A*f*GT Consortium:

> *is a valid method for assessing whether a teacher's performance meets the Australian Professional Standards for Teachers at the Graduate Teacher level. The panel noted that this is a very well designed and executed project, delivered at a relatively early stage of maturity. It is believed that further reliability and validity data, and analysis of cross-institutional similarities and differences will strengthen this TPA as time goes on…. The expert panel endorses the AfGT as meeting the requirements of Program Standard 1.2 at this point in time. Given the data limitations at this stage of the instrument's development, the panel recommends that areas 'in progress' should be brought back to the panel for reconsideration in twelve months' time (AITSL EAG, May 2018).*

Twelve months later, the Expert Advisory Group provided the following advice:

> *The expert panel found that the AfGT is reflected in a well thought out and thorough TPA that demonstrates a valid and reliable method for assessing whether a teacher's performance meets the Australian Professional Standards for Teachers at the Graduate Teacher level. The panel noted that in this resubmission, the AfGT and the University of Melbourne have provided clear responses to the advice from the EAG 12 months ago, as well as its own learning through its implementation process. The panel noted that consistent monitoring of how the TPA is implemented across providers, along with changes to assessment processes may need to occur in the future if inconsistencies in applying the TPA occur (10 July, 2019).*

This report details progress since the last report to the Consortium dated 8 August 2020 and includes ongoing technical analysis of data related to the implementation of the instrument as well as developments within the Consortium itself.

Image by Monoar Rahman Rony from Pixabay

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

2. Consortium Update

2.1 Consortium Members

The following institutions are Consortium Collaborators in the A*f*GT Consortium:

- University of Melbourne (Lead Institution)
- Charles Darwin University
- Curtin University
- Federation University
- University of Canberra
- University of Sydney
- University of Western Australia
- University of Technology Sydney
- Victoria University

The following institutions are Consortium Licensees:

- Montessori Institute, Western Australia (commenced early 2019)
- Excelsia College, Sydney (commenced early 2020)
- Melbourne Polytechnic (commenced mid 2021)
- Southern Cross Education Institute (commencing start 2022)
- University of Adelaide (commencing start 2022)

2.2 Response to COVID-19

Full implementation in all programs in all the member institutions was anticipated to occur during 2020, however the COVID-19 crisis affected schools, particularly those in Victoria, in a number of ways. The following figure provides a summary of the periods of remote and flexible learning that interspersed periods of face-to-face teaching in Victoria as well as the requirements of RISEC, the Victorian Department of Education and Training's Research in Schools and Early Childhood Settings. Victoria's situation is highlighted here because of its extended period of lockdown from 9 July to 27 October 2020, which was reportedly amongst the most severe and heavily policed periods of lockdown in the world (BBC News, 2020).

The Executive Group of the A*f*GT Consortium established a COVID-19 Response Team (CRT) to assist institutions in their implementation of the A*f*GT during and following the COVID-19 crisis. The CRT met on a fortnightly basis, and more frequently as required. Early in April 2020, the Consortium Chair, Professor Janet Clinton, reassured the consortium that:

> the A*f*GT can and should continue to be implemented with ***fidelity*** as required by the National Expert Group, and I realise that may present some challenges. We must, of course, be mindful of not risking our PSTs' registration opportunities, and most importantly, add to the risk to our programs. It will be important to consider that the A*f*GT was not designed to be a stand-alone assessment and that is embedded in our ITE programs. In saying this, I understand that you are facing different circumstances and processes in variable contexts, and that each institution will need to make their own decision about their programs to align with their institution's guidelines (Clinton, 2020).

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

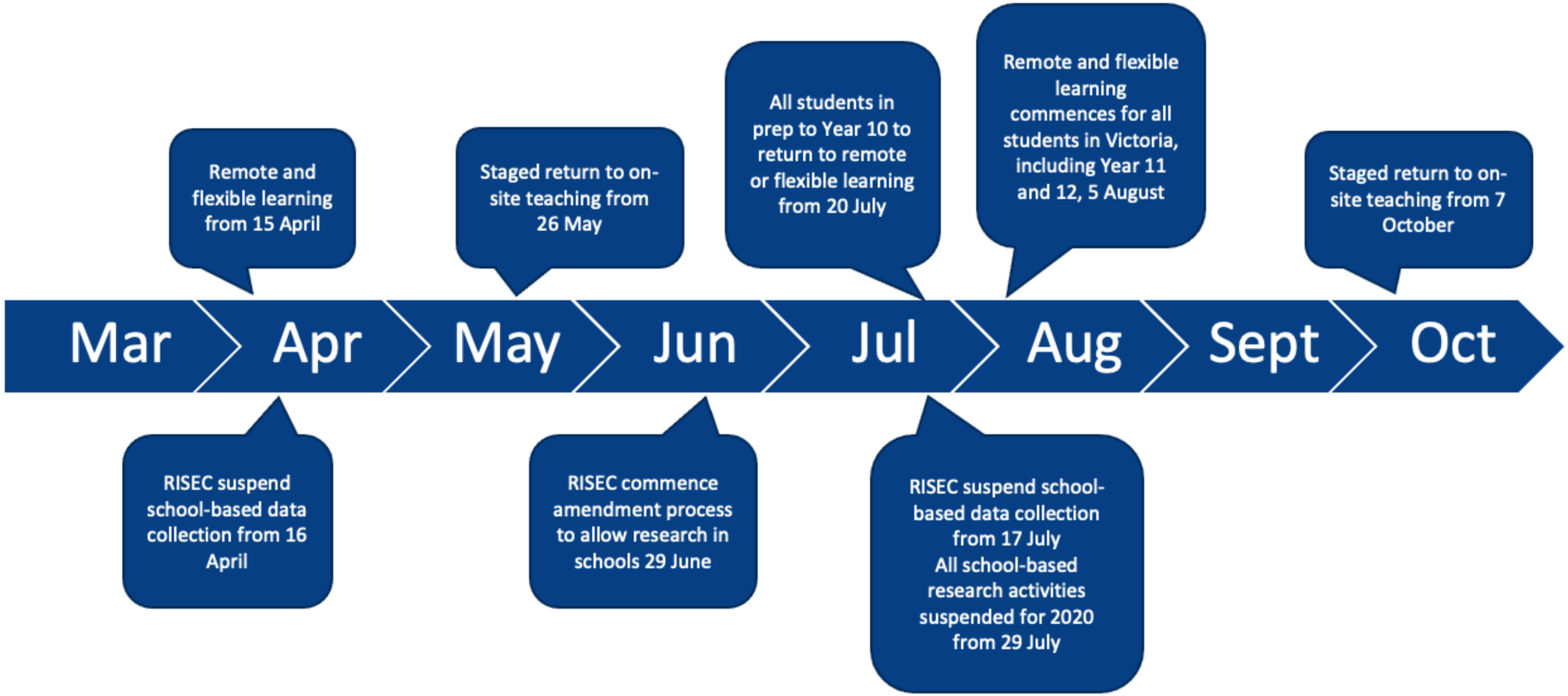


Figure 1. 2020 Timeline of COVID-19 disruptions to teaching, learning and research activities

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

The CRT's Terms of Reference specified the four roles of the team, namely:

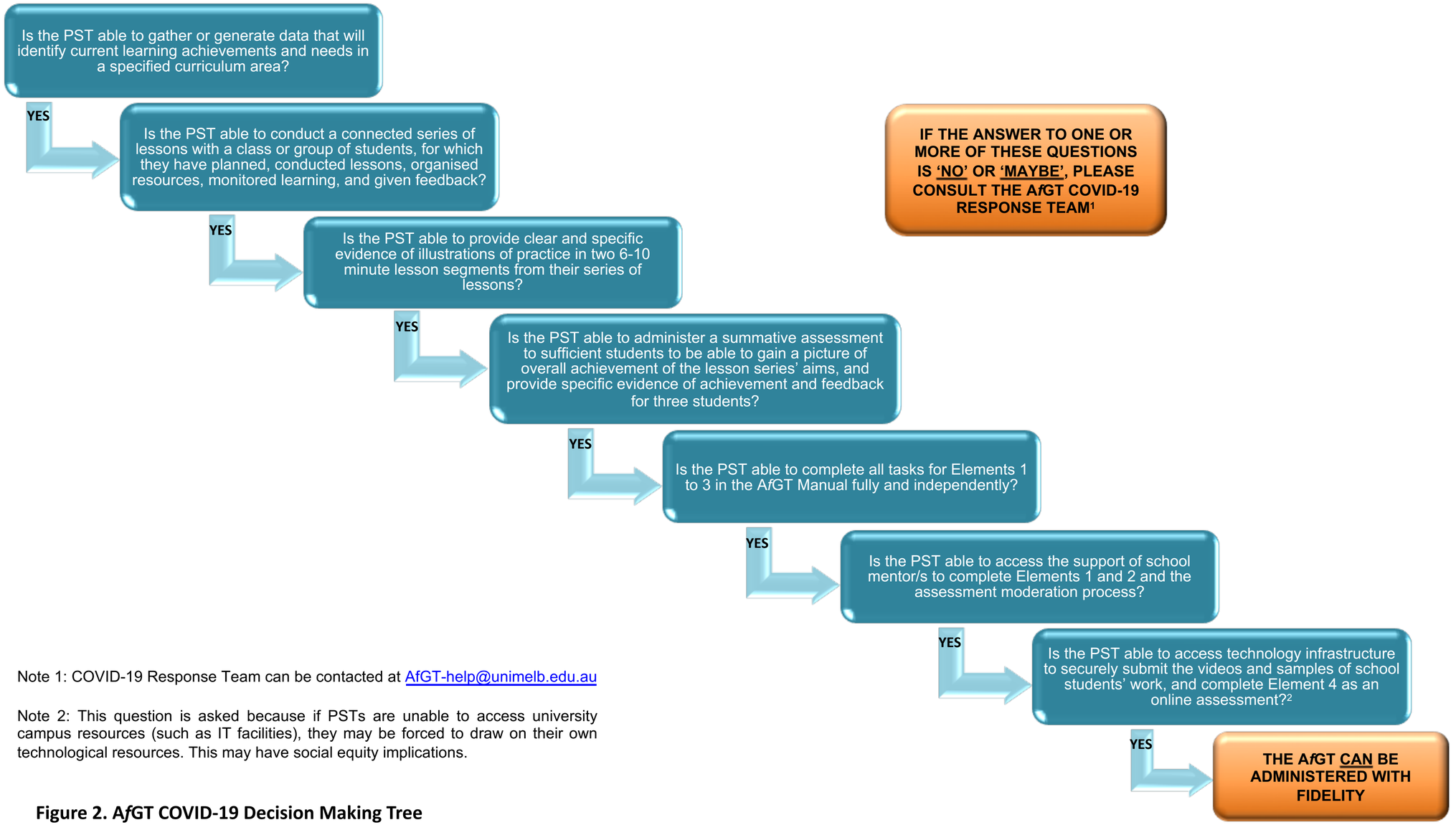
- Guidance Role: to provide notes and FAQs regarding adaptations to maintain the fidelity of the instrument whilst cohering to jurisdictional guidelines,
- Consultative Role: to provide feedback to institutions on a case-by-case basis,
- Oversight Role: to review adaptation strategies, including advising Executive should adaptations fall outside of the scope of the instrument, and a
- Monitoring and Reporting Role: to work with the A*f*GT Management Team to:
    - Provide updates on the latest policy positions
    - Review register of issues and solutions
    - Collect data about strategies that institutions used to cope with COVID-19
    - Create a platform for institutions to share resources
    - Monitor and report to the Consortium how institutions have implemented the A*f*GT with fidelity, and
    - Canvass the latest updates from institutions and jurisdictions and make recommendations to the Executive for communication required with various stakeholders.

To assist institutions to make decisions about the adaptations that they might be considering, the CRT developed two key decision-making tools: the A*f*GT Decision Making Tree (see Figure 2) (and associated guidance package) and a series of matrices in which scenarios were 'tested' and shared using the decision-making tree (refer Table 1).

The A*f*GT's member-only website (housed on the University of Melbourne's Learning Management System) was also expanded to provide guidelines and announcements from external stakeholders, scholarly articles and practical suggestions around remote and online learning, including resources generated by teacher educators from within the Consortium. Despite the interruptions caused by the pandemic, 2348 PSTs from eleven institutions undertook the A*f*GT in 2020.

Photo by Jessica Ruscello on Unsplash

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References



Note 1: COVID-19 Response Team can be contacted at AfGT-help@unimelb.edu.au

Note 2: This question is asked because if PSTs are unable to access university campus resources (such as IT facilities), they may be forced to draw on their own technological resources. This may have social equity implications.

Figure 2. A*f*GT COVID-19 Decision Making Tree

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

Hypothetical Scenarios	Is the PST able to gather or generate data that will identify current learning achievements and needs in a specified curriculum area?	Is the PST able to conduct a connected series of lessons with a class or group of students, for which they have planned, conducted lessons, organised resources, monitored learning, and given feedback?	Is the PST able to provide clear and specific evidence of illustrations of practice in two 6 to 10-minute lesson segments from their series of lessons?	Is the PST able to administer a summative assessment to sufficient students to be able to gain a picture of overall achievement of the lesson series' aims, and provide specific evidence of achievement and feedback for three students?	Is the PST able to complete all tasks for Elements 1 to 3 in the AfGT Manual fully and independently?	Is the PST able to adequately access the support of school mentor/s to complete Elements 1 and 2 and the assessment moderation process in Element 3?	Is the PST able to access technology infrastructure to securely submit the videos and samples of school students' work, and complete Element 4 as an online assessment?[1]	**Can the AfGT be administered?**
3. The Education Department advises (prior to the commencement of the placement) that the placement must be completed in a shorter timeframe than the intended time period for the placement. Does this adaptation meet the requirements of the AfGT?	YES	YES	YES	YES	MAYBE	MAYBE	MAYBE	YES, as long as the placement meets accreditation, institutional course requirements and the adjustments were approved by the jurisdiction's Teacher Registration Board (TRB) AND that the PSTs are able to complete a sequence of between 5 and 8 lessons. Arrangements may need to be made in order for the PST to submit video segments, students' assessment tasks and complete Element 4.

Table 1. A*f*GT COVID-19 Decision Making Matrix

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

2.3 Jurisdictional and Regulatory Responses

Table 2. Summary of jurisdictional and regulatory responses to COVID-19

Summary of Jurisdictional Temporary Regulatory Measures in Response to the COVID-19 Pandemic	
Australian Children's Education & Care Quality Authority (ACECQA)	***March 26, 2020:*** • Acknowledgement that theoretical aspects of programs can be delivered online, but that supervised professional experience may not be possible at this time. ***May 7, 2020, with reminder on August 4, 2020:*** • Undergraduate programs to include at least 30 days of supervised professional experience in EC settings, including min. 10 days with children birth to 35 months. Postgraduate programs to include at least 20 days in EC settings, including min. 10 days with children birth to 35 months. • Providers expected to facilitate "other meaningful forms of technology and scenario-based assessment, such as tele-presence, simulations and work-integrated placements" where traditional placements are not possible.
New South Wales Education Standards Authority (NESA)	***May 8, 2020:*** • Providers should negotiate with sectors, schools and centres, localised arrangements for the placement of ITE students to assist in the delivery of teaching and learning through online/remote teaching modes as well as alternative learning opportunities. • Providers should endeavour to maintain as far as possible the minimum accreditation standard for graduating students having completed at least 60 or 80 days of professional experience. • Providers, in consultation with schools and centres, should try to maximise the amount of face-to-face teaching for individual ITE students. • NESA's overriding expectation of providers is that the assessment of final year students through the above mix of online/remote teaching and face-to-face teaching and alternative learning opportunities and their Teaching Performance Assessment continues to be based on demonstrating the necessary Graduate Teacher Standards to the satisfaction of supervising teachers and provider staff, rather than the precise number of days of professional experience completed.
Teacher Registration Board of Western Australia (TRBWA)	***April 8, 2020:*** • "The minimum number of professional experience days complete by any pre-service teacher enrolled in an accredited ITE program, and due to complete in 2020, should comprise at least 45 days." • Providers are expected to ensure all final year PSTs complete at least 25 of the days in 2020.
Victorian Institute of Teaching (VIT)	***April 23, 2020:*** • Minimum number of professional experience days reduced to at least 60 days (reduced from 80 days) for undergraduate and 45 days (reduced from 60 days) for graduate ITE programs, including professional experience undertaken online. • ITE providers must be able to declare it is sufficiently assured the PST has met all the Graduate Teacher Standards. • It is expected that PSTs successfully complete TPA. • Providers required to explain to VIT how supervised teaching practice is realised in an online learning context. ***September 18, 2020:*** • Revised measures largely reiterated advice from April. • Minimum number of professional experiences days "by any PST enrolled in an accredited ITE program due to complete in 2020 or mid-year in 2021 will be reduced to 45 days", including professional experience undertaken online or as part of Victorian DET's Small Group Tutoring initiative. VIT encouraged providers, where possible, to exceed the revised baseline of 45 days. • Supervised teaching practice – either in a school or non-school setting – must be supervised by a registered teacher or a person able to be registered as a teacher. • Measure extended to PSTs completing combined early childhood/primary ITE programs requiring PSTs to have completed the majority of 45 placement days to be completed in a school setting (primary context). Programs must also meet ACECQA's minimum requirements.

2.4 AfGT Sustainability Plan

In addition to providing a response to COVID-19, members of the Executive have been formulating a draft AfGT Sustainability Plan, which it is anticipated will be completed in October 2021. The purpose of the Sustainability Plan is to formalise the underpinning assumptions that inform the Consortium's activities.

2.5 Governance Arrangements

All institutions have signed the Collaboration Agreement 2019 – 2024, which has guided the activities of the Consortium. The governance structure (Figure 3) has remained stable, as have the number of committees and their functions.

During 2020, member feedback suggested that there be more interaction between committee leads and the Executive Group. Consequently, several changes were made to the way that committee leads have interacted (formally) with members of the Executive Group:

- Committee leads and co-leads joined with the Executive Group for a Committee Kick-off Meeting on 17 February 2021, to provide an update of activities and to discuss work plans for 2021.
- Executive meeting schedules now include a timetable so that the leads and co-leads from each committee on a rotational basis meet with the Executive at the start of each Executive Group meeting. The purpose of these meetings is to provide focused updates of work undertaken and to discuss any issues for discussion/clarification that have been identified by the committee.

Photo by Elisa Calvet B. on Unsplash

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

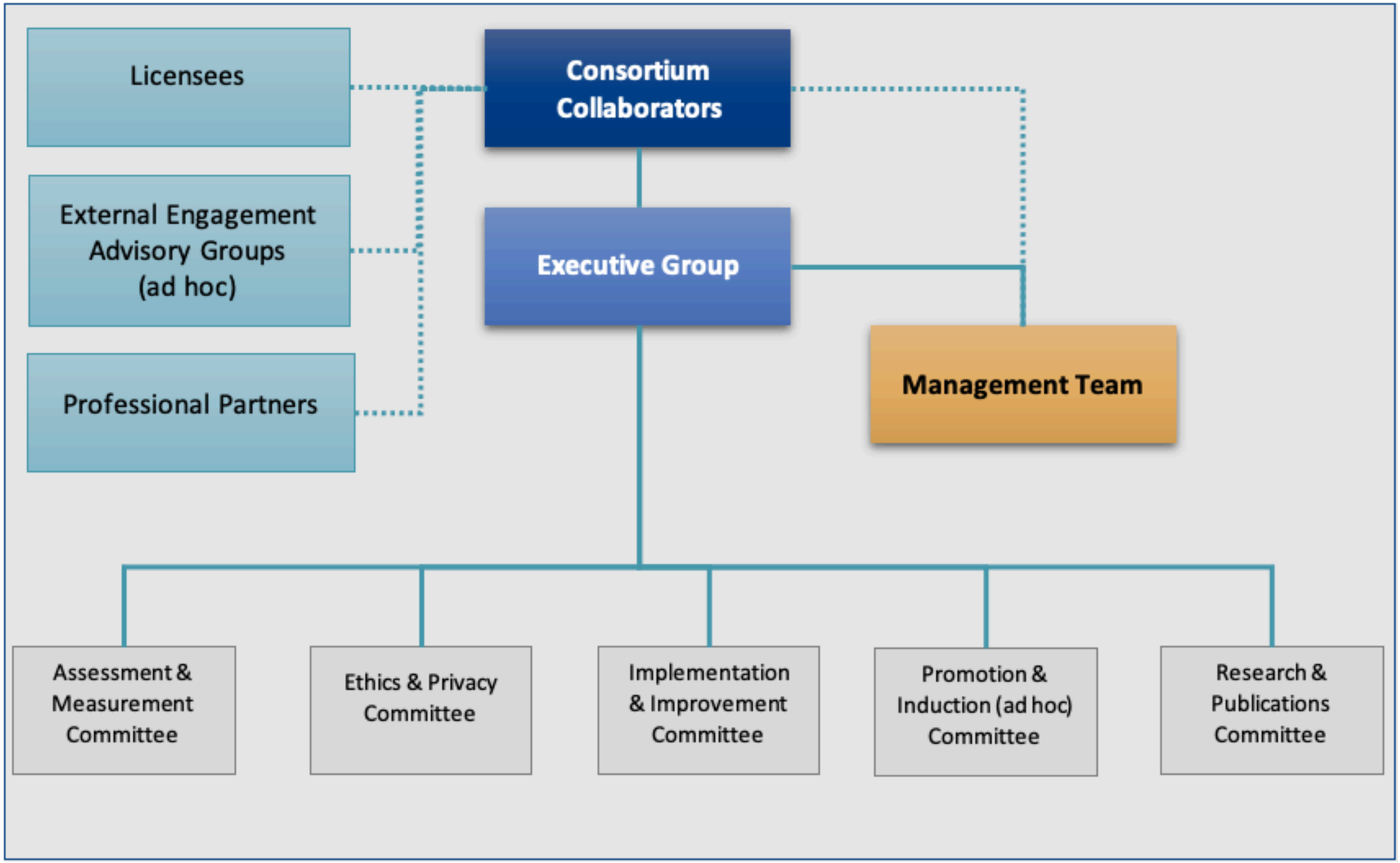


Figure 3. A*f*GT Consortium governance structure

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

2.6 Activities of the Consortium and Executive Group

Members of the Consortium are actively involved in the running of the consortium. The arrangements in 2020 for meetings of different entities or groups within the Consortium were as follows:

Table 3. Entities Within the A*f*GT and Meeting Frequency

Entity	Meeting Frequency
Consortium Collaborators & Licensees	Quarterly, and as required
Executive Group	Monthly, and as required
COVID-19 Response Team	Fortnightly, and as required, reviewed at the end of each 3-month period
Committees	Minimum of 4 meetings per year
Deans/Heads of Schools	Twice yearly
A*f*GT Management Team	Weekly, and as required

The table overleaf summarises the main points reported to the Executive Group at the meeting on 17 February 2021.

Photo by Hannah Busing on Unsplash

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

Table 4. Committee achievements since last report

	Committee Achievements
Assessment & Measurement Committee (AMC)	a.Positive feedback was received following the 2020 online moderation workshops, b.An email encouraging institutions to circulate invitations to survey and focus group meetings to evaluate A*f*GT processes, though data was thought to be very limited for 2020 due to the interruptions caused by COVID, c.The key tasks for the AMC in 2021 are: • Supporting moderation exercises by developing guidelines/practices for each institution to create a systematic view on how moderation works (or will work) at each institution, • Analysis of the moderation data, • Committee intends to start inter-rater reliability assessments, by mid-2021, once they have obtained 3 sets of moderation data from institutions, • Create a feedback/note space for assessors to document and share new situational judgments, and • AMC agrees with IIC that more fine-grained work (i.e. alignment between the task and the assessment rubrics) needs to be conducted on the A*f*GT instrument.
Ethics & Privacy Committee (EPC)	a.The most significant accomplishment for EPC for 2020 was the creation of A*f*GT's Privacy Statement, including infographics and communication slides, b.Currently, ethics approvals across various jurisdiction are ongoing and being negotiated, c.The key tasks for the EPC in 2021 are: • Ensuring ethics amendments mean that Consortium members/institutions can share data at conferences and publish in journals, • Submit ethics amendments to the UoM Ethics Committee and other jurisdictional bodies, and • Providing examples of how the privacy statement can be linked back to course requirements.
Implementation & Improvement Committee (IIC)	a.Last year's work was primarily focussed on: • Development of new Element 4 scenarios along with a process for trialling and review, and • Consideration of processes to enable timely feedback re continuous improvement of the instrument. b.The key tasks for the IIC in 2021 are: • Compilation of a living register where assessors can record areas where alignment between the task and the assessment rubrics could be improved, • Ensuring recommended changes are in place in readiness for the first cohorts in 2022, and • Development of support materials targeting school-based staff re implementation of the A*f*GT.
Research & Publications Committee (RPC)	a.Last year's work was primarily focused on continuing collaborative writing of articles and clarifying the data to which Consortium members have access, b.The key tasks for the RPC in 2021 are: • Exploring ethics requirements for using de-identified scripts used in the moderation workshops in publications, and • Determine publishing priorities following contact being made by a book publisher expressing interest/opportunity to submit a book for publication.
Promotion & Induction Committee (PIC) ad hoc	The work that was intended for this committee, mainly in relation to the induction of new licensees, has been undertaken by the Director, and Project Manager, A*f*GT Management Team.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

2.7 Benchmarking Project (AITSL)

Following a recommendation from the Education Council (June 2018) that AITSL should lead a benchmarking exercise with all approved TPAs to confirm the passing standard and to confirm that the TPAs were assessing PSTs' competence against the APSTs consistently, the A*f*GT Management Team participated in several meetings with AITSL and the other two consortia—the Graduate Teaching Performance Assessment (GTPA) and the Quality Teaching Performance Assessment (QTPA)—in October to contribute to the design of the benchmarking activity. Towards the end of 2020, this approach to benchmarking was abandoned, partially due to increasing numbers of single institutions having their TPAs approved by the Expert Advisory Group.

Subsequent to this, Professor Janet Clinton received a formal request for a proposal to submit a TPA cross-institutional moderation research paper—due September 30, 2021—to:

- explore the essential elements of moderation and cross-institutional moderation (CIM),
- build capacity and provide professional learning for those involved in TPA development and implementation, and
- support the implementation of consistent and rigorous cross-institutional moderation processes across all TPAs and hence provide further assurance that all TPAs are valid and reliable.

2.8 Publications/Conferences

The Research & Publications Committee has developed documentation to record the planned and completed conference presentations and publications. These documents are located on the Consortium's LMS and are available for all members to access and update. Details of publications, conference presentations and submissions for conferences are included below.

2.8.1 Publications since last report

Keamy, R. K., & Selkrig, M. A. (2021). Interrupting practice traditions: Using readers' theatre to show the impact of a nationally mandated assessment task on initial teacher educators' work. *Teaching Education*. https://doi.org/10.1080/10476210.2021.1951198

Kriewaldt, J., Walker, R., Morey, V. Morrison, C. (2021) Activating and reinforcing graduates' capabilities: Early lessons learned from a Teaching Performance Assessment. *Australian Education Researcher*. https://doi.org/10.1007/s13384-020-00418-4

McGraw, A., Keamy, R. K., Kriewaldt, J., Brandenburg, R., Walker, R., & Crane, N. (2021). Collaboratively designing a national, mandated teaching performance assessment in a multi-university consortium: Leadership, dispositions and tensions. *Australian Journal of Teacher Education*. http://dx.doi.org/10.14221/ajte.2021v46n5.3

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

2.8.2 2021 Conferences

European Educational Research Association, Geneva, 2021 (Sept 6-10: online)

Paper title: Partnering to build a teaching performance assessment: Perspectives on designing a national, mandated assessment instrument in a cross-university collaboration. (Accepted for the 2020 EERA Conference and re-submitted and accepted for the 2021 conference.)

Collaborators: Amanda McGraw, Robyn Brandenburg, Nadine Crane, Rebecca Walker, Jeana Kriewaldt, Kim Keamy

Paper presented by Jeana Kriewaldt and Nadine Crane on 8 September 2021.

AARE Conference, Melbourne, 2021 (Nov 28-Dec 2: online)

1. ***Symposium title:*** The Tudge Review: How the A*f*GT teaching performance assessment represents professionalism during rapid policy churn and frequent review of Initial Teacher Education

Paper titles:

i. Discerning key principles for a nationally mandated teacher performance assessment: Literature Review for A*f*GT.
ii. Perspectives on designing a national, mandated assessment instrument in a cross-university collaboration: Considering some of the social costs and benefits.
iii. The impact of a nationally mandated assessment task on initial teacher educators' work using readers' theatre: Act 2.
iv. Activating and reinforcing graduate capabilities: Early lessons learned from a Teaching Performance Assessment.

Symposium Collaborators: Janet Clinton (chair), Diane Mayer (discussant), Rebecca Walker, Robyn Brandenburg, Jeana Kriewaldt, Kim Keamy, Nadine Crane, Amanda McGraw, Mark Selkrig, Valerie Morey.

2. ***Paper title:*** Determining the Sustainability of Teacher Performance Assessments

Collaborators: Janet Clinton, Kim Keamy, Val Morey, Wayne Cotton, Emily Hills, Katina Tan

Photo by Markus Spiske on Unsplash

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

3. Findings From 2020 Data

3.1 About the A*f*GT instrument

The A*f*GT is made up of four elements, each with a different number of items (tasks and sub-tasks). A series of rubrics accompany each element of the A*f*GT. These rubrics are criterion-referenced on a developmental continuum so that performance can be assessed through successive levels of increased competence.

Indicative behaviours described at each level of the rubrics have been developed with the integrated use of taxonomies such as Blooms', Krathwohl's, SOLO and Dreyfus' model of skill acquisition. The rubric levels range from 1 through to 4, with 4 being the highest achievement.

- Level 4 (G+) indicates the PST exceeds the Graduate Standard,
- Level 3 (G) indicates the PST is at the Graduate Standard,
- Level 2 (G-) indicates the PST is not yet at the Graduate Standard, and
- Level 1 (U) represents that there is insufficient and/or unsatisfactory information in the response for a judgement to be made.

The indicative behaviours at Level 3 are calibrated to the relevant Graduate Teacher Standard.

The PST **is required to pass all four elements** of the A*f*GT to demonstrate that the Australian Professional Standards for Graduate Teachers (AITSL, 2018) are met. However, it is possible that not all tasks within an element are passed, which means that the assessor will need to make an 'on balance' judgement whether each element has been passed. Meeting the Graduate Teacher Standards enables the PST to graduate from the respective accredited programs of learning. The PST cannot graduate unless they have satisfactorily completed the A*f*GT and all other course assessment tasks.

Figure 4 provides a summary of the four inter-related assessment elements.

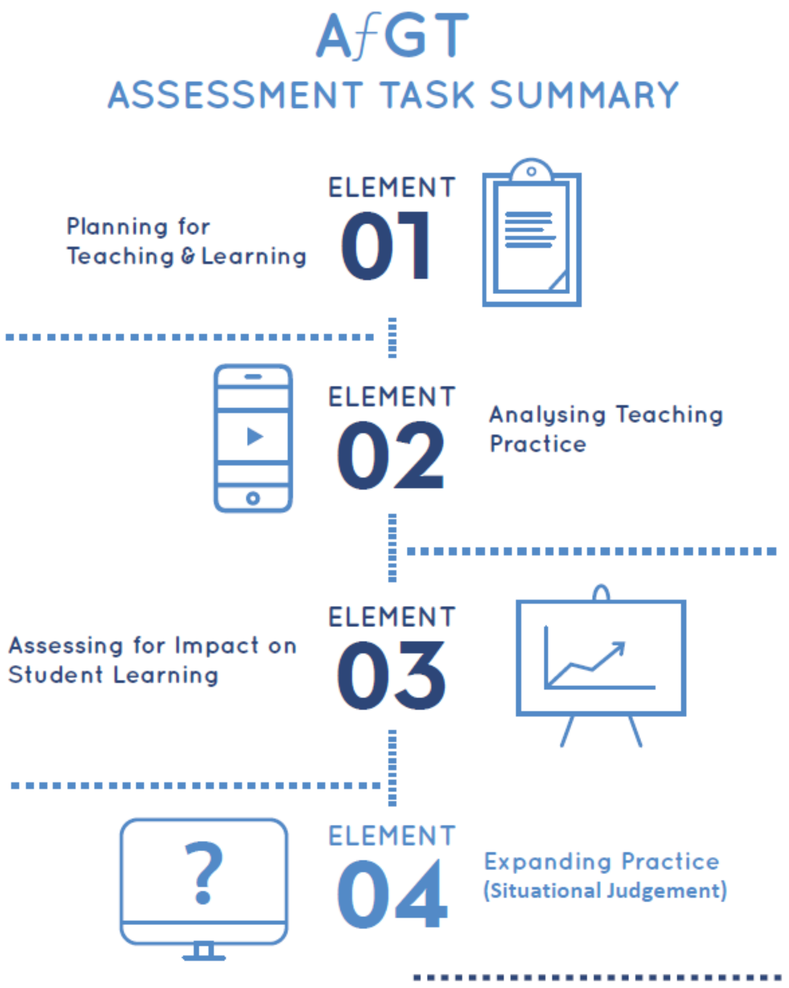


Figure 4. A*f*GT assessments task summary

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

3.2 Findings

Overall, 2348 PSTs completed the A*f*GT across eleven institutions in 2020. Consistent with prior years, there were significantly more female PSTs (65%) in the 2020 cohort compared to males (30%) and other genders. The breakdown between undergraduate and postgraduate PSTs, which is determined by the programs offered by the respective institutions, was almost equal between masters (47%) and bachelor programs (53%). Within the bachelors, the largest cohort was the Bachelor Primary (24%), whereas the largest Masters cohort was the Masters Secondary (36%). A more comprehensive breakdown of the participants is presented in Table 5. The sample appears to fit the profile of PSTs across the Consortium and is considered representative.

Table 5. Participant Demographics

	2019 *(%)*	2020 *(%)*
Gender		
Female	1255 *(75%)*	1528 *(65%)*
Male	421 *(25%)*	697 *(30%)*
Other	-	1 (0%)
Missing Data	-	122 (5%)
Program Type		
Bachelor Early Childhood	96 *(6%)*	107 *(5%)*
Bachelor Primary	373 *(22%)*	552 *(24%)*
Bachelor Secondary	268 *(16%)*	282 *(12%)*
Bachelor EC/Primary	48 *(3%)*	27 *(1%)*
Bachelor Primary/Secondary	-	267 *(11%)*
Masters Early Childhood	-	12 *(0%)*
Masters Primary	179 *(11%)*	207 *(9%)*
Masters Secondary	646 *(39%)*	857 *(36%)*
Masters EC/Primary	66 *(3%)*	37*(2%)*
TOTAL	**1676**	**2348**

**Missing data denotes no information provided for the gender variable*

As described previously, the A*f*GT comprises four elements, each containing several interrelated tasks, as shown in Figure 4. To understand the average and distribution of participants' scores, the mean and standard deviation of scores for each element were calculated and are presented in Table 6. Distribution of grades across the sample are presented in Figure 5.

Table 6. Mean scores and SD by element

	n	Average score	Std. deviation
Element 1	2004	3.34	0.37
Element 2	1994	3.23	0.37
Element 3	1987	3.25	0.38
Element 4	2294	3.20	0.42

The histograms in Figure 5 shows the distribution for the four A*f*GT elements with the vertical lines at scores of 2, 2.5 and 3. A score of 2 represents a grade of 'G-', whilst a score of 3 represents a grade of 'G'. For illustration, a vertical line of 2.5 is also represented, as this is the nominated cut-score for the Consortium.

Collectively, these results demonstrate high consistency in the distribution of grades across the four elements, suggesting consistency in scoring across the elements. The data also suggest that there is an opportunity for success in each of the elements and a relatively equivalent score across the elements. This is significant, and provides support that the instrument is highly stable, given the varying number of tasks (or items) in each element.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

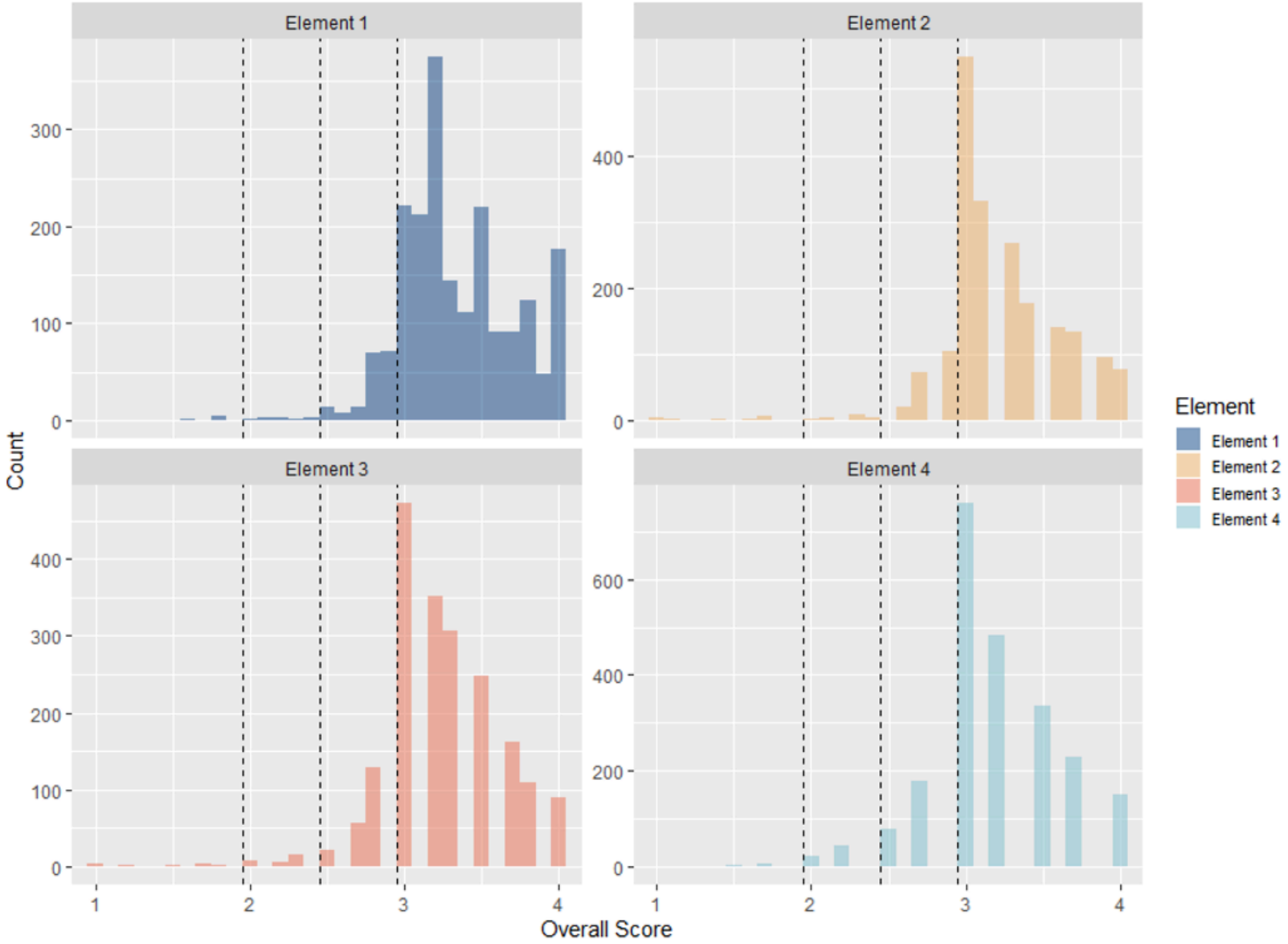


Figure 5. Grade distribution by element

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

3.2.1 Results by Institution

A summary of results by institutions is presented in Table 7 and the mean score by element for each institution is presented in Figure 6. Out of the eleven institutions, eight had a sample size that was larger than 50 PSTs, and these were reported as discrete institutions. The other three institutions with smaller cohorts of less than 50 PSTs are reported as a combined 'Others' group.

Table 7. Mean Scores and SD by Each Element by Institutions

	n	Element 1	Element 2	Element 3	Element 4
Institution A	494	3.20 *(0.20)*	3.22 *(0.29)*	3.22 *(0.26)*	3.23 *(0.29)*
Institution B	456	3.19 *(0.40)*	3.12 *(0.41)*	3.16 *(0.45)*	2.88 *(0.48)*
Institution C	419	3.45 *(0.35)*	3.30 *(0.47)*	3.40 *(0.46)*	3.32 *(0.48)*
Institution D	343	3.35 *(0.27)*	3.30 *(0.32)*	3.28 *(0.34)*	3.31 *(0.33)*
Institution E	204	3.34 *(0.28)*	3.18 *(0.33)*	3.17 *(0.35)*	3.26 *(0.29)*
Institution F	190	3.95 *(0.15)*	3.29 *(0.39)*	3.36 *(0.38)*	3.17 *(0.35)*
Institution G	152	3.40 *(0.38)*	3.37 *(0.40)*	3.42 *(0.39)*	3.29 *(0.37)*
Institution H	57	3.09 *(0.19)*	3.07 *(0.30)*	3.04 *(0.46)*	3.23 *(0.44)*
Others	33	3.07 *(0.39)*	3.11 *(0.42)*	3.03 *(0.35)*	3.34 *(0.37)*
Total	**2348**	**3.34 *(0.37)***	**3.23 *(0.37)***	**3.25 *(0.38)***	**3.20 *(0.42)***

With the exception of Element 1 for Institution F, the distribution of grades for each element across each institution is fairly consistent, although there is some variability between each element for Institutions B, F and H. For Institution F, initial investigation suggests that the distribution may have been impacted by the assessment policy regarding resubmissions for the institution. Nonetheless, further analysis and discussion with will be required before any conclusions can be drawn. As expected, the variability in the 'Others' category is higher than the other institutions, given the smaller sample size and the scores were sourced from three different institutions.

Institutions that have offered their data for analysis will receive confidential individual reports that relate to their own institution in comparison to the overall Consortium sample data. These individual institution reports will be circulated separate to this report.

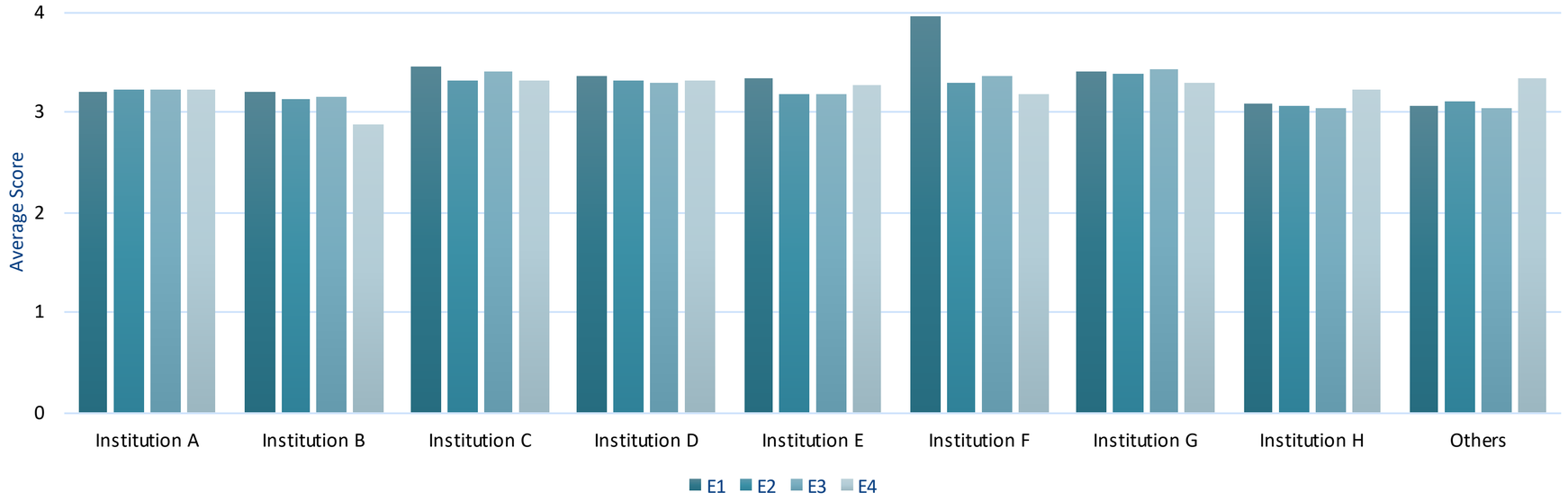


Figure 6. Average scores by each element by institution

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

3.3 Scale Reliability and Instrument Validation

This section presents the reliability analysis and instrument validation for A*f*GT using factor analysis, Item Response Theory (IRT) analysis and Differential Item Functioning (DIF) analysis. In examining the validity of the instrument, the 2020 data was combined with 2019 data given that no changes were made to the instrument. This provides a more comprehensive analysis and reflects the cumulative nature of the data that provides validity evidence for the instrument.

3.3.1 Factor structure and Internal Coherence

Exploratory Factor Analysis (EFA) was conducted to investigate whether the items making up the four elements of the A*f*GT group together as theorised, indicating that the four elements represent independent factors that measure unique aspects of classroom readiness.

Maximum Likelihood factor extraction with an oblimin (oblique) rotation was used to find the most parsimonious factor solution, as the factors are expected to be correlated. Participants with incomplete data were removed from the factor analysis, leaving 3478 participants. Overall, the results of factor analysis and reliability estimates remained consistent with 2019 in terms of how items were clustered and the reliability values for each scale, providing support that the instrument remains highly valid and reliable.

A plot of the correlation matrix (shown in Figure 7) also seems to suggest a two, three, four, or five factor solution.

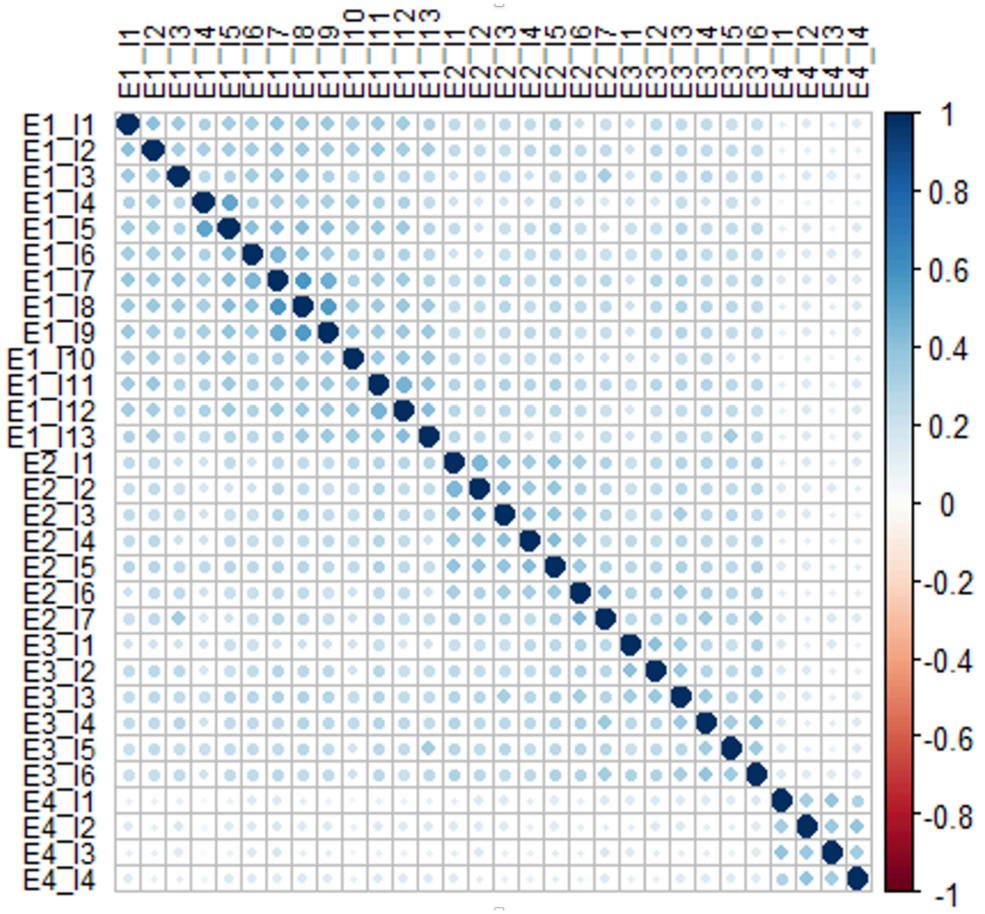


Figure 7. Correlation matrix of A*f*GT items

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

Upon further analysis, the three or four factor solutions are likely to be the best fit for the data. Table 8 shows the three and four factor solutions with the respective Cronbach alpha measures whilst Figure 8 shows the visual representation of the three and four factor solutions respectively.

Table 8. Factor Solution with Estimate of Reliability

	Items	Cronbach's α	
Three-factor Structure			
Factor 1	Element 1 items	0.873	**Very good**
Factor 2	Element 2 and 3 items	0.849	**Very good**
Factor 3	Element 4 items	0.679	**Moderate**
Four-factor Structure			
Factor 1	Element 1 items	0.873	**Very good**
Factor 2	Element 2 items*	0.763	**Good**
Factor 3	Element 3 items*	0.784	**Good**
Factor 4	Element 4 items	0.679	**Moderate**

Under the three-factor structure, Element 1 and Element 4 are unique factors, while Element 2 and Element 3 group together to measure a single construct. Under the four-factor solution, all four elements were separated into unique factors, although two items from Element 2 loaded on the "Element 3" factor. In both cases, the Cronbach's alpha measures indicated that the items on the scale displayed "Moderate" to "Very good" reliability. The "Moderate" scale for both the three-factor and four-factor structures applied to Element 4. This may be impacted by the number of items that was lower than those for the other factors. Collectively, these results seem to indicate that items in Element 1 and Element 4 separate clearly from the rest of the items, while Element 2 and Element 3 items are marginally closer related to each other.

Given that an exploratory factor analysis methodology was employed, without pre-determining the number of factors, this is an encouraging result which confirms the internal coherence of the instrument. The four-factor structure in particular, coheres well with the overall design of the instrument. As the data sample expands in the future, it could be possible that the factor analysis may be further refined to detect a clearer separation between Element 2 and Element 3 items. It may also be worthwhile to further investigate the Element 2 items that are loading on to Element 3. Notwithstanding, ongoing validation will continue by using a consistent methodology of factor analysis to ensure that the factor structure of the instrument remains robust and psychometrically defensible in the future datasets.

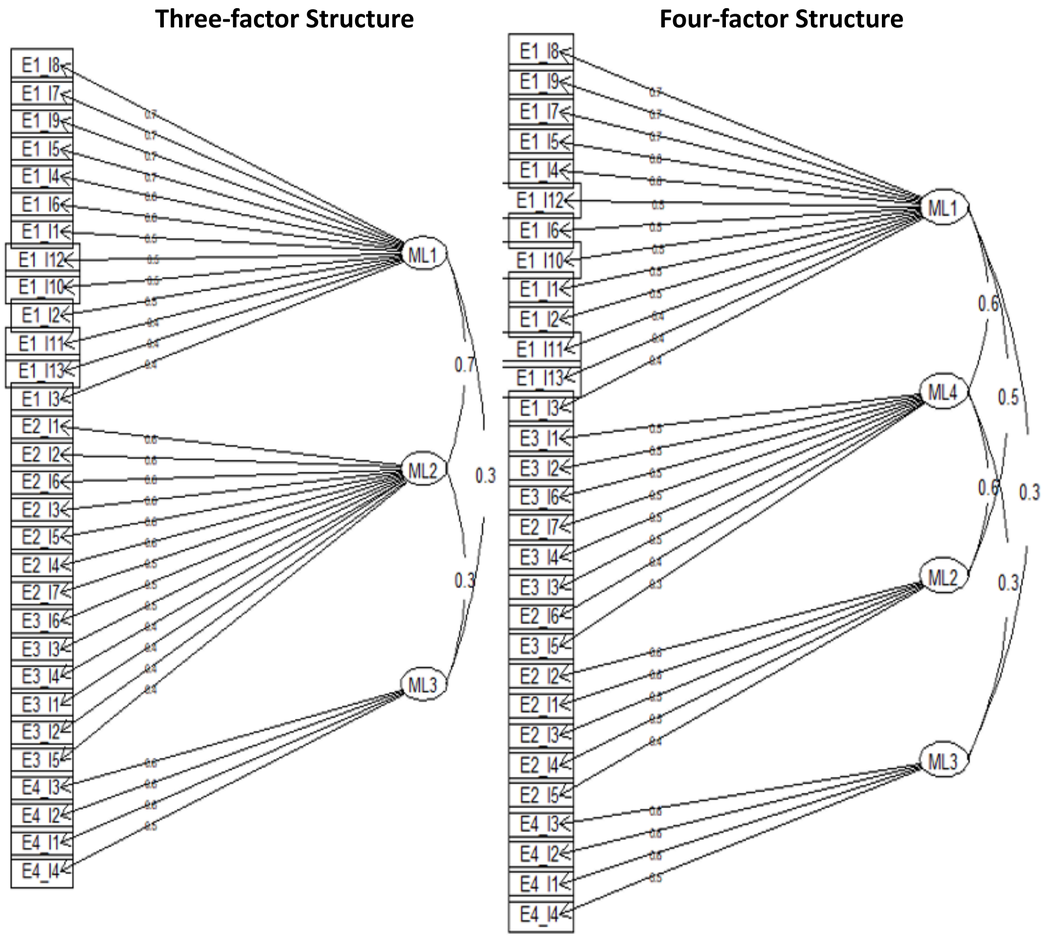


Figure 8. Factor analysis using three-factor and four-factor structures

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

3.3.2 Item Analysis

In adhering to the framework for establishing A*f*GT's assessment validity and reliability shown in Figure 26, Item Response Theory (IRT) analysis was employed to analyse the A*f*GT items. This section discusses the IRT analysis in terms of model fit, item statistics, item ordering and test information from the A*f*GT data.

1) Model Goodness-of-Fit

For the purposes of the A*f*GT Item Response Theory (IRT) analysis, two models were considered applicable to the current data: the Graded Response Model (GRM) and the Generalised Partial Credit Model (GPCM). These two models were selected because they can be applied to polytomous items that use an ordered response scale, and they provide informative discrimination and item difficulty parameters.

Other IRT models such as the one-parameter logistics model (1PLM) and two-parameter logistics model (2PLM) were excluded as these models are for dichotomous items. Models that have a guessing parameter were also excluded as it is generally not possible for PSTs to 'guess' the correct answer for A*f*GT, given the nature of the assessment. Furthermore, models such as the Partial Credit Model (PCM) and the Rating Scale Model (RSM) were also deemed less suitable as they assume equal discriminability across all items and the RSM estimates a single set of categorical location parameters for all items, making these models less informative than the GRM and GPCM models for the A*f*GT data (Muraki, 1992; Nguyen, et al., 2014; Zanon et al., 2016).

To gauge how well the two chosen models can predict PSTs' scores and generate item statistics that are invariant over the data set, a comparison of model fit for both GRM and GPCM was performed. Table 9 shows the goodness-of-fit statistics of the two models. As expected, both models provided very similar results, with the GRM model having slightly better model fit statistics. This is further supported by the model comparison statistics shown where both AIC and BIC values are lower for the GRM model and likelihood ratio tests were significant ($p < 0.001$). As such, it is deemed more useful to employ the Graded Response Model (GRM) and the results in the next few sections are reported based on the four-dimensional GRM model output.

Table 9. Goodness-of-Fit Statistics

Model	χ^2	df	RMSEA	SRMSR	TLI	CFI	AIC	BIC
GPCM	-	345	0.0504	0.1851	0.9099	0.9171	150251.0	150989.5
GRM	768.789	345	0.0561	0.1844	0.8883	0.8972	149482.2	150220.7

2) Item Discrimination and Item Difficulty Parameters

Table 10 represents the estimated item-level discrimination parameters ***('a')*** and item-level difficulty parameters ***('b')*** for each item of the A*f*GT instrument.

The discrimination parameter indicates how well an item distinguishes between PSTs who display different levels of classroom readiness. As a rule of thumb, values >1 indicate good discriminability. While very high values can occur, they may indicate a problem with the assessment. For this reason, values between 1 and 4 are generally seen as ideal. All the items in the A*f*GT meet this criterion and confirms the notion that the instrument can effectively distinguish PSTs who are at different levels of ability.

The ***('b')*** parameter displays the threshold position on the z-distribution of the latent construct (in this case, the latent construct is teaching readiness) between two levels on the response scale. As the A*f*GT items are scored on a four point scale, there are three ***('b')*** parameters for each item (indicating the threshold between 'U' and 'G-', 'G-' and 'G', and 'G' and 'G+').

For example, a b_1 value equalling -4.61 for Item 1 of Element 1 indicates that PSTs who are 4.61 standard deviations (SDs) below the mean score would be expected to score a 'U' on this item, whereas those that are above this value will score a 'G-', up until they reach 2.742 SDs below the mean (as $b_2 = -2.742$), at which point they would be expected to score a 'G', and so on.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

Based on this analysis, the most difficult item to achieve a 'G', which indicates a PST is at Graduate Standard of the relevant Australian Professional Standards for Teachers, is Element 3 Item 6 which requires PSTs to justify the next steps in their teaching based on an evaluation of assessment data and an understanding of research into how students learn. The easiest item to achieve a 'G' is Element 2 Item 6 where PSTs are required to evaluate the adjustments that have been made to their teaching based on observation, evidence and mentor feedback.

The most difficult item to achieve a 'G+' which indicates that a PST exceeds the Graduate Standard of the relevant APST is Element 1 Item 10, where PSTs are required to synthesize their mentor's feedback to support their planned learning sequence. And the easiest item to achieve a 'G+' is Element 1 Item 6 where PSTs are required to design sequenced lesson content that includes curriculum links.

Table 10. Item Statistics

	Discrimination Parameter (a)	Difficulty Thresholds (b)		
		b_1	b_2	b_3
Element 1				
Item 1	1.782	-4.610	-2.742	0.340
Item 2	1.750	-3.126	-2.180	0.542
Item 3	1.599	-3.806	-2.715	-0.358
Item 4	1.506	-3.813	-2.328	0.656
Item 5	2.081	-3.737	-2.482	0.476
Item 6	2.135	-3.348	-2.504	-0.439
Item 7	2.608	-3.155	-2.157	-0.015
Item 8	2.732	-3.120	-2.089	0.130
Item 9	2.204	-2.899	-1.876	0.514
Item 10	1.714	-3.356	-1.921	1.058
Item 11	1.703	-4.122	-2.242	0.534
Item 12	1.897	-3.715	-2.228	0.869
Item 13	1.553	-3.026	-1.932	0.958
Element 2				
Item 1	2.017	-3.351	-2.123	0.966
Item 2	1.942	-3.478	-2.191	0.807
Item 3	2.021	-3.200	-2.170	0.650
Item 4	1.862	-3.059	-2.147	0.885
Item 5	1.955	-3.052	-2.039	0.800
Item 6	1.580	-4.215	-2.861	0.230
Item 7	1.269	-3.231	-2.593	0.712
Element 3				
Item 1	1.454	-4.327	-2.596	0.236
Item 2	1.633	-4.133	-2.493	0.543
Item 3	1.739	-4.069	-2.618	0.263
Item 4	1.657	-3.703	-2.065	0.807
Item 5	1.314	-3.577	-2.281	0.945
Item 6	1.677	-3.430	-1.826	0.890
Element 4				
Item 1	1.534	-4.548	-2.054	0.649
Item 2	1.707	-4.262	-2.131	0.559
Item 3	1.789	-3.732	-2.201	0.493
Item 4	1.547	-4.615	-2.339	0.658

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

3) Item Ordering

Using the same information computed for item difficulty thresholds, it is also possible to analyse the probabilities of the rating categories ('U', 'G-', 'G', and 'G+') for each item to ensure that the ordered categories are correctly distinguishing the abilities of PSTs. This is best depicted using a graphical representation known as a Category Characteristic Curve (CCC) (see figures 9-12). In a CCC, the x-axis represents the z-distribution of a PST's classroom readiness based on the entire sample, where θ=0 is the mean, and the range of -6 to 6 representing standard deviations (SD). The y-axis represents the probability of receiving a given grade, from a range of 0 (0% chance) to 1.0 (100% chance).

Each line therefore, represents the probability of receiving a grade for a given ability level depicted on the x-axis. For example, the blue line in Figure 9 represents the probability of receiving a 'U' grade from 6 SDs below the mean all the way to 6 SDs above the mean. As expected, the lower the PST is on the ability level, the higher the chance of receiving a 'U' grade as it is the lowest possible grade that can be awarded. It is worth noting here that whilst the y-axis range is 0 to 1.0, the probability curves are asymptotic, i.e. they would only approach the extreme, but would not actually be 0 or 1.0. This is because it is impossible to say that we can be 100% sure a PST will get a 'U' grade, no matter how low the PST's ability level (Wilson, 2005).

The points where the adjacent categories intersect represent transitions from one category to the next. Specifically, the point where two curves intersect is the point on the z-distribution of the PSTs' ability where there is an equal (0.5) probability of being awarded either of the two grades (e.g. the intersection point between 'U' and 'G-' indicates that PSTs with that level of classroom readiness have a 50% chance of being awarded a 'U' or a 'G-'.)

Thus, using the same example of Element 1 Item 1 (Figure 9, item label 'E1_I1' most bottom left corner), PSTs who are 4.61 SDs below the mean score would be expected to score a 'U' and this is depicted by the intersection of the blue and pink dotted line. PSTs who score above this value are expected to get a grade level of 'G-' until they reach 2.742 SDs below the mean, where the dotted pink line intersects with the dotted green line. And PSTs above this value are expected to get a grade level of 'G' until they reach 0.34 SDs above the mean, where the dotted green line intersects with the dotted red line. PSTs who are above 0.34 SDs above the mean are expected to get a grade level of 'G+'.

The CCC visualisation is an easy way to identify if there are any disordered categories, where, for example, the intersection of the dotted pink and dotted green line appears to the left of the blue and dotted pink intersection. Step disordering would mean that there is a probability that a PST who performs lower than a peer may be awarded a higher grade than the higher-ability peer. Figures 9 to 12 depicts the CCC for each A*f*GT item by element.

Except for Element 2 Item 7 which appears to separate participants into three levels rather than the expected four, the CCC curves show that all the A*f*GT items are correctly ordered. Element 2 Item 7 requires PSTs to synthesise research into how students learn to justify adjustments made to their teaching practice. For this item, there appears to be a merging of grade levels 'U' and 'G-'. This suggests that the Consortium may want to consider further refining the rubric for this item such that there is a clearer distinction between grade level 'G-' and grade level 'U'.

It is also worth noting that the A*f*GT items are more effective in separating PSTs at the low end of the classroom readiness scale (x-axis), and less effective in separating PSTs who display higher than mean levels of classroom readiness (the 0 point on the x-axis). This is reflected by most of the inflection points (b_1, b_2 and some b_3) being located below the mean in Table 10. This is further discussed in the next section.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

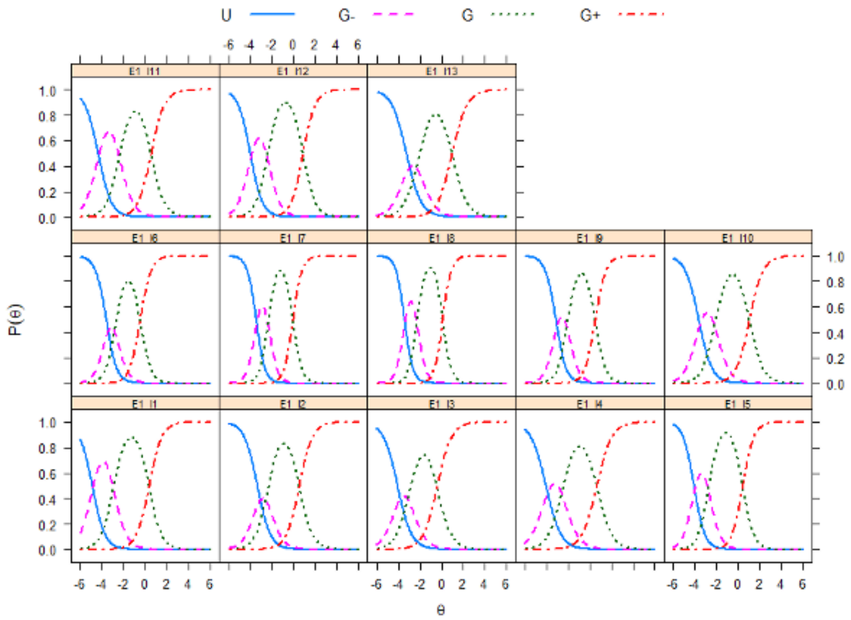


Figure 9. Category Characteristics Curves for Element 1

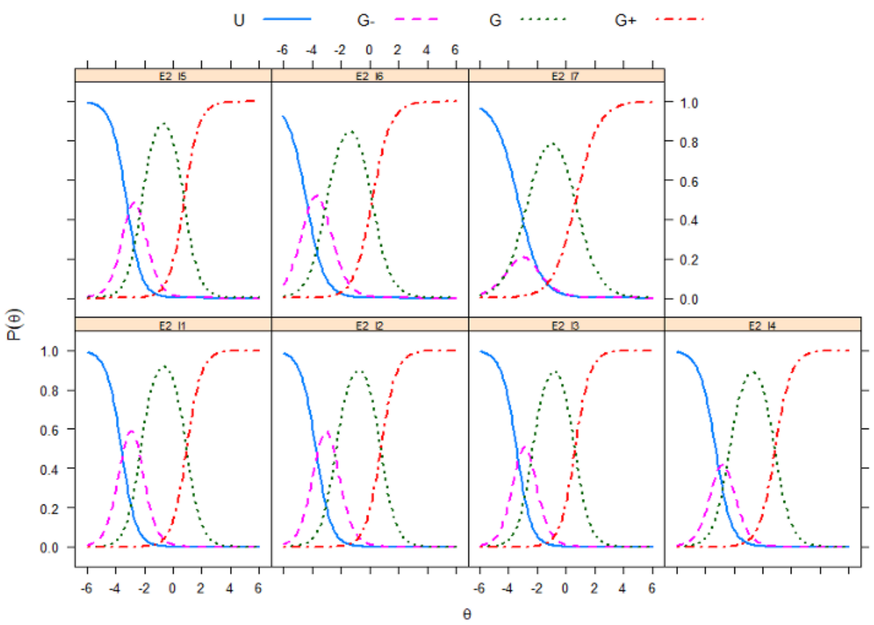


Figure 10. Category Characteristics Curves for Element 2

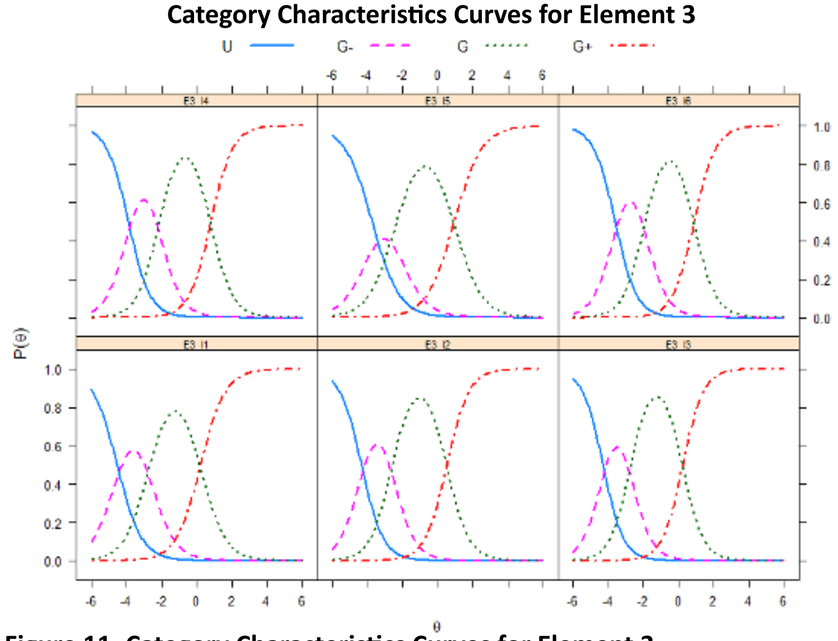


Figure 11. Category Characteristics Curves for Element 3

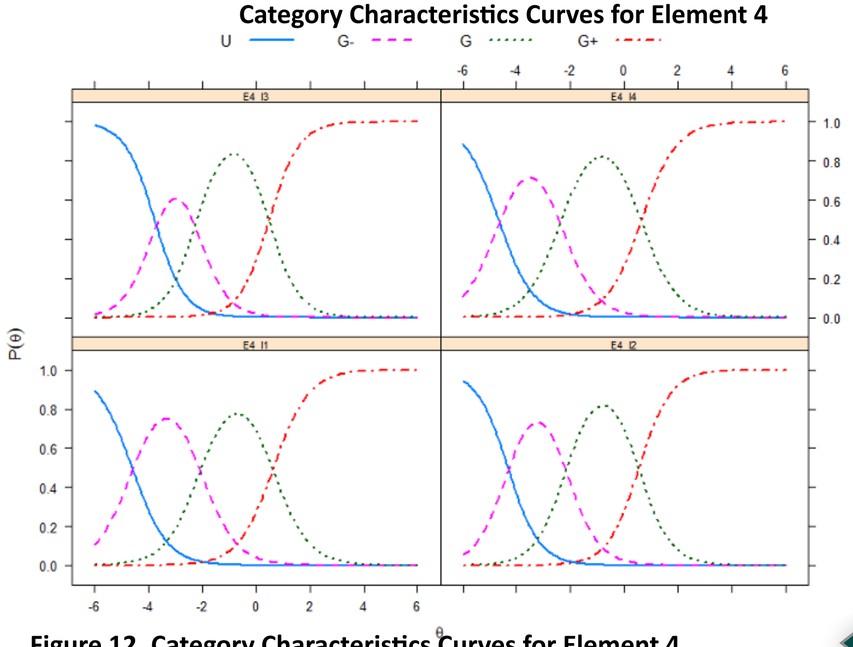


Figure 12. Category Characteristics Curves for Element 4

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

4) Test Information

Figure 13 overleaf shows the Test Information Function (TIF) for the four elements of A*f*GT. TIF plots indicate how well an instrument estimates PSTs' location on the classroom readiness scale (x-axis). The y-axis represents the amount of information at its estimated difficulty parameter. Information curves indicate the points on the teaching readiness continuum where the test is best able to distinguish between students.

For example, for Element 1, the test provides maximum information for PSTs located at the lower end of the scale (left side of x-axis); there is then a drop before a small peak for PSTs approximately located at θ=0, followed by a steep drop in the test information on the higher end of the scale (right side of x-axis). This suggests that the A*f*GT is very effective at obtaining precise estimates of a PST's readiness to teach if they are between 2 and 4 SDs below θ. It is also effective at estimating PSTs at the mean score. However, the A*f*GT is not effective at providing precise estimates for PSTs at the higher end of the score scale. This is consistent for all four elements of the A*f*GT.

Given that the A*f*GT is not a ranked assessment, but rather a criterion-based assessment that focuses on PSTs' classroom readiness, the TIF is consistent with the conceptual design of the instrument. For all four elements, the A*f*GT is highly effective at obtaining precise estimates of PSTs who are 'on-the-cusp' (θ=0) where it is critical to determine if a PST has indeed met the APST at Graduate level. Element 4 reflect this particularly well. In the future, it may be useful to consider refining the instrument such that Elements 1, 2 and 3 are also able to provide maximum information for PSTs approximately located at and around the mean.

5) Element 4 Chi-square and Item Analysis

Separate to the analyses performed for the overall instrument, Element 4 was given particular focus as this element consists of multiple scenarios (or items) for each item set. This is consistent with the Consortium's strategy to develop an item bank for Element 4 to minimise plagiarism and ensure validity of the instrument.

Because PSTs are randomly assigned one of six possible scenarios for each item set, it is important to ensure that there is no internal bias between the scenarios. To do this, Chi-square and item analyses was employed, along with an investigation of the descriptive statistics of the data. In this analysis, the 2020 dataset was used which includes 2129 participants after removing incomplete data. Overall, there is empirical evidence to support the fairness claim across the item sets, and across the scenarios within each set.

The Chi-square analysis (which is a test of independence) found no evidence of dependency between the allocation of scenario and the performance of the PSTs. In other words, whether PSTs pass or fail Element 4, was not significantly dependent on which scenarios they were assigned. Furthermore, the item analysis found no substantial difference in the difficulty levels among the scenarios. Both findings suggest that there is no internal bias among the Element 4 scenarios.

The item analysis also found all the scenarios having good item statistics, suggesting that none of them need to be discarded. There was one scenario within the first item set that was slightly underfitting, but still within an acceptable range. This suggests a closer review of the item may be beneficial, to check for clarity and coherence. Due to the confidential nature of the Element 4 scenarios, the full analysis is not provided in this report. However, a technical report relating to this section is available to Consortium members upon request.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

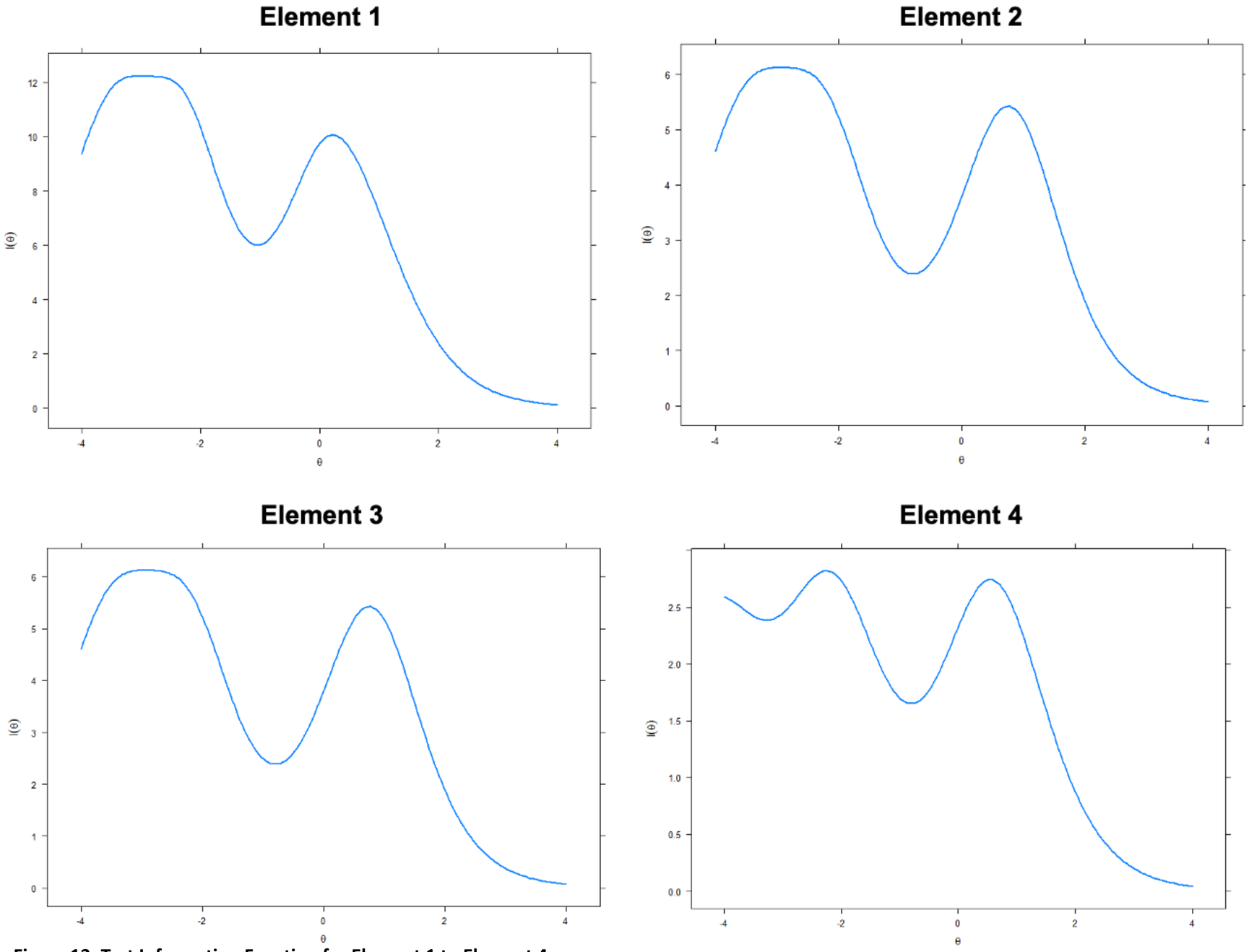


Figure 13. Test Information Function for Element 1 to Element 4

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

3.3.3 Fairness Evidence

This section investigates the probability of PSTs receiving different scores on the A*f*GT based on the group they belong (bachelor, masters, primary, secondary, early childhood, etc.) rather than based purely on their level of classroom readiness. Analysis is performed at two levels; first a descriptive summary and an analysis of difference is determined for each group. Where differences are significant, a Differential Item Functioning (DIF) analysis is conducted to analyse the source of difference. DIF is a statistical method that detects differences in the probability of obtaining a grade level ('U', 'G-', 'G', and 'G+') for each subgroup (Acar, 2011).

1) Analysis of bachelor vs masters

Table 11 presents a descriptive summary of the four elements for Bachelor and Masters program. Across the four elements, the mean difference between Bachelor and Masters ranged from -0.04 to 0.04 (out of the maximum of 4). To determine whether two groups are statistically different from each other, a t-test is performed between Bachelor and Masters PSTs. Element 2 and Element 3 did not show a significant difference at the 0.05 level.

Element 1 and Element 4 showed a significant difference at p < .001. However, this outcome could be due to the large samples (large enough to detect a small but significant difference by program). Given the effect size or magnitude of difference indicated by Hedges' g is very small (0.125 and 0.109 for Element 1 and Element 4 respectively), it is possible to conclude that the A*f*GT does not bias PSTs based on their grouping of Bachelor and Masters program.

Table 11. Descriptive Summary by Element for Bachelor and Masters program

	n	Element 1 Mean *(SD)*	Element 2 Mean *(SD)*	Element 3 Mean *(SD)*	Element 4 Mean *(SD)*
Bachelor	1848	3.35 *(0.34)*	3.26 *(0.35)*	3.27 *(0.36)*	3.23 *(0.41)*
Masters	1630	3.31 *(0.36)*	3.24 *(0.40)*	3.27 *(0.41)*	3.29 *(0.44)*
T-Test		*t*(3476) = 3.691, *p* < .001	*Not significant*	*Not significant*	*t*(3476) = 3.221, *p* = .001
Hedges' *('g')*		0.125	N/A	N/A	0.109
Total	**3478**				

2) Analysis of program type

Program type is comprised of five groups; Secondary, Primary, Early Childhood, Primary/Secondary Combined ('Pri/Sec') and Early Childhood/Primary Combined ('EC/Pri'). Table 12 presents a descriptive summary of these five groups based on the four A*f*GT elements.

To determine whether three or more groups are statistically different from each other, an analysis of variance (ANOVA) is performed by element. For all four elements, ANOVA was significant (p < .001), suggesting that program type makes a difference to the scores.

Table 12. Descriptive Summary by Element for program type

	n	Element 1 Mean *(SD)*	Element 2 Mean *(SD)*	Element 3 Mean *(SD)*	Element 4 Mean *(SD)*
Secondary	1721	3.32 *(0.35)*	3.24 *(0.39)*	3.28 *(0.40)*	3.30 *(0.42)*
Primary	1181	3.37 *(0.34)*	3.30 *(0.36)*	3.30 *(0.37)*	3.30 *(0.42)*
Early childhood	210	3.34 *(0.30)*	3.23 *(0.32)*	3.26 *(0.27)*	3.31 *(0.35)*
Pri/Sec	205	3.18 *(0.39)*	3.08 *(0.31)*	3.10 *(0.41)*	2.71 *(0.49)*
EC/Pri	148	3.41 *(0.42)*	3.17 *(0.39)*	3.20 *(0.37)*	3.35 *(0.36)*
ANOVA		F(4, 3460) = 16.30, p < .001	F(4, 3460) = 17.28, p < .001	F(4, 3460) = 13.53, p < .001	F(4, 3460) = 107.94, p < .001
Total	**3465**				

To further understand the differences by program type, Differential Item Functioning (DIF) analysis was conducted for Primary vs Secondary PSTs. Because DIF is influenced by sample size, Early Childhood, Pri/Sec and EC/Pri specialisation have been excluded from the analysis. Primary and Secondary program type are the two largest groups in the sample, and initial analysis is showing differences in scores between these two groups in Element 1 and Element 3. The DIF analysis provides information on the extent to which Primary and Secondary program PSTs who are equal in terms of classroom readiness display different results for A*f*GT items.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

An ordinal logit-model based DIF was conducted using Bonferroni-corrected likelihood ratio tests to identify group-level differences in either discrimination ***('a')*** or difficulty ***('b')*** parameters for each of the A*f*GT items. Data were standardised (z-scored) before entry into the model. Category Characteristic Curves (CCC) for Primary and Secondary PSTs were plotted for each item. The output of the analysis is presented in Table 13. Four items were detected as DIF items:

- Element 1 Item 5
- Element 1 Item 11
- Element 3 Item 2
- Element 3 Item 3.

Of these, Element 1 Item 5 was significant at the p<.001 level and is investigated further below.

To better understand the difference between Primary and Secondary PSTs, a Category Characteristic Curve (CCC) was plotted for Element 1 Item 5. CCC presents an easy visualisation of the DIF analysis outcome as shown in Figure 14 overleaf. For each of the coloured curves (the colour represents grade levels), there are two sets of lines representing the two groups of PSTs analysed. The black curves represents 'U' grade, yellow represents 'G-' grade, blue represents 'G' grade and pink represents 'G+'. The solid line denotes Secondary PSTs and the dash-dotted line denotes Primary PSTs.

For this item, there is no clear indication of which group received higher scores, in other words this item is not biased towards Primary or Secondary PSTs. However, at grade 'G', there is a wider score distribution for Secondary PSTs compared to Primary PSTs. This is represented by a broader and flatter blue solid curve compared to the dash-dotted blue curve. Secondary PSTs also received better scores than Primary PSTs at the 'G+' level, represented by the solid pink curve located to the right of the dash-dotted pink curve. Interestingly, this is reversed at the lower grade levels of 'U' and 'G-', where Primary PSTs are receiving better scores than Secondary PSTs (the black and yellow dash-dotted curves are located to the right of the solid black and solid yellow curves respectively).

Table 13. DIF output for Primary vs Secondary program type

	Chi-square value	p value	Adjusted p value	**
Element 1				
Item 1	2.248	0.325	1.000	
Item 2	5.007	0.082	1.000	
Item 3	9.217	0.010	0.299	
Item 4	0.323	0.851	1.000	
Item 5	27.692	0.000	0.000	***
Item 6	4.642	0.098	1.000	
Item 7	6.153	0.046	1.000	
Item 8	11.604	0.003	0.091	*
Item 9	0.958	0.619	1.000	
Item 10	5.069	0.079	1.000	
Item 11	14.713	0.001	0.019	**
Item 12	8.608	0.014	0.405	
Item 13	11.413	0.003	0.100	*
Element 2				
Item 1	8.760	0.013	0.376	
Item 2	2.895	0.235	1.000	
Item 3	2.821	0.244	1.000	
Item 4	8.526	0.014	0.422	
Item 5	2.835	0.242	1.000	
Item 6	0.687	0.709	1.000	
Item 7	4.651	0.098	1.000	
Element 3				
Item 1	8.113	0.017	0.519	
Item 2	14.777	0.001	0.019	**
Item 3	12.996	0.002	0.045	**
Item 4	5.894	0.053	1.000	
Item 5	4.503	0.105	1.000	
Item 6	1.244	0.537	1.000	
Element 4				
Item 1	0.056	0.972	1.000	
Item 2	2.394	0.302	1.000	
Item 3	1.776	0.411	1.000	
Item 4	9.685	0.008	0.237	

Significant codes:
** p < .1*
*** p < .01*
**** p < .001*

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

This item requires PSTs to design a cumulative sequence of lesson goals that align with the overarching goal. A possible factor that may be affecting the DIF for this item could be due to the bigger variability in the types of lesson plans that are assessed for Secondary PSTs, resulting in higher variability of scores. If the same pattern emerges for the DIF results in future years, then there is a possibility that a systematic inconsistency may be occurring between Primary and Secondary PSTs undertaking the A*f*GT.

To address this, the Consortium could consider focusing its moderation efforts on this area, both at institution and cross-institution level. For example, a cross-section of Element 1 Primary and Secondary scripts could be selected for moderation to align assessors' view on equivalence between the different types of lesson plans and sequence of lessons assessed across the Primary and Secondary program types.

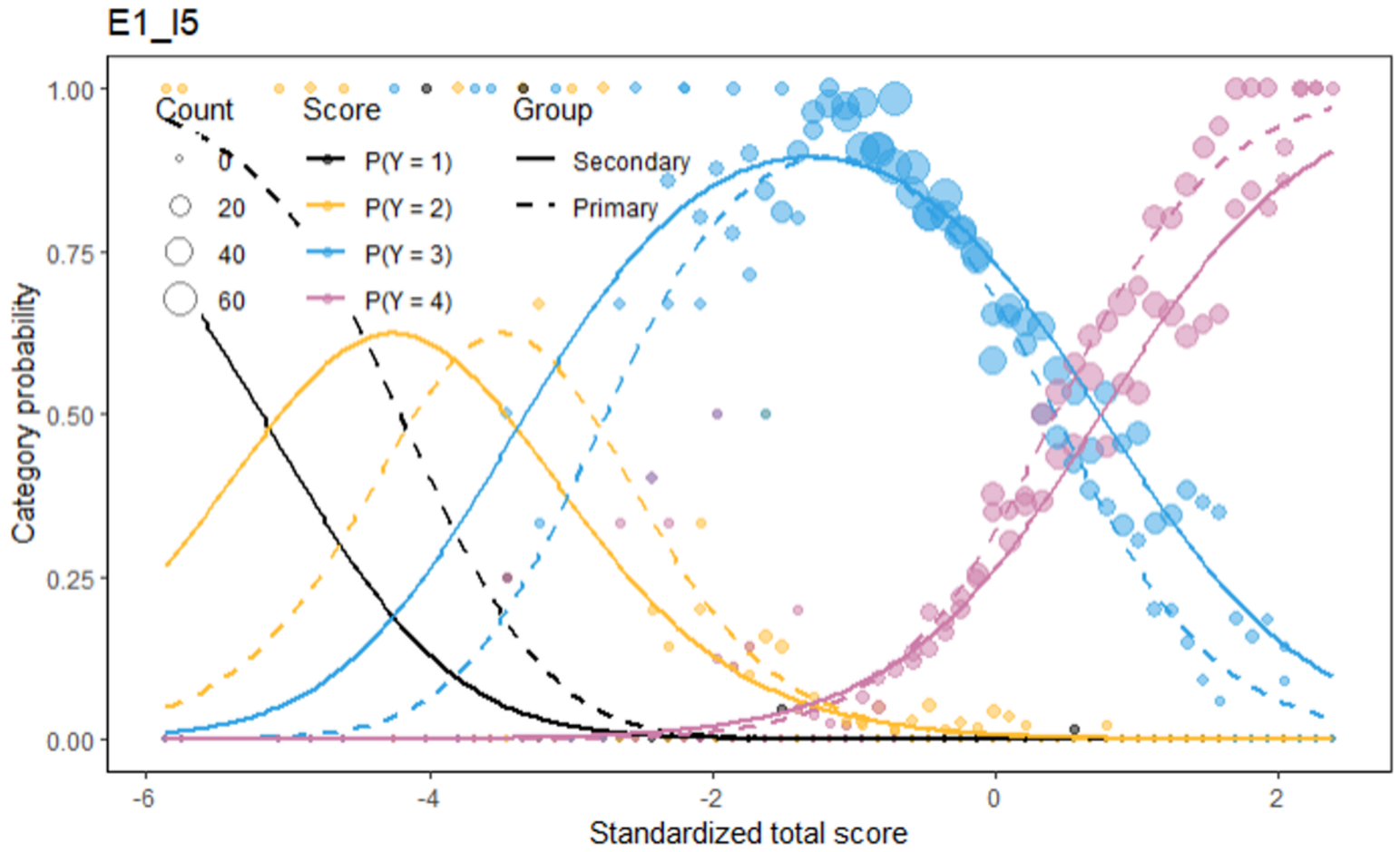


Figure 14. Category Characteristic Curve for Element 1 Item 5 (Primary vs Secondary)

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

3.4 Validation of the Cut Score

In meeting Program Standard 1.2 (AITSL, 2019) on Teaching Performance Assessments, the TPA instrument must have a clear, measurable and justifiable achievement criteria that discriminate between meeting and not meeting Graduate Teacher Standards. AITSL further expands on this requirement to include:

- *evidence that the standard for successfully completing the TPA is set at a level that reflects the Graduate Teacher Standards*
- *a credible process for differentiating those who meet the standard and those who do not.*

As the TPA is a high stakes assessment, the tools and processes used to determine the criteria for meeting and not meeting the standard are crucial. For example, evidence needs to reference the use of a recognised professional standard setting methodology to determine the passing threshold.

The A*f*GT Consortium uses several methods based on empirical evidence to ensure the methodology employed to determine its cut score continues to be valid and reliable. This includes synthesising the findings from factor and item analysis, cumulative performance data and coherence with the conceptual framework of the A*f*GT instrument design. To differentiate PSTs who meet the standard and those who do not, a cut score is applied at level 'G', which is defined as '*meeting APST standards at graduate level*'. Achieving a level 'G' is deemed the required level to pass each element. To pass the A*f*GT, PSTs are required to pass all four elements.

In applying a conceptual approach to calculating the cut score for level 'G', the analysis was broken down by element, given that PSTs had to pass all four elements. The raw grades, 'U', 'G-', 'G', 'G+' were converted to 1, 2, 3, 4, respectively. Then, for each element, based on the number of tasks, all possible combinations to obtain a score between '2' and '3' was identified. The mean for each element is calculated and the overall cut score is the mean of the four elements. Using this method, the calculated cut score is **2.57**.

A post-hoc analysis is conducted by applying the calculated cut score to the 2020 data sample, the findings of which are presented in Table 14. As with previous years, the overall pass/fail distribution reflects a lower proportion of PSTs passing the A*f*GT (91%) as compared to the proportion of passes in each element (between 93% to 99%), due to the requirement of passing all four elements. As the distribution data are consistent with previous years, the analysis suggests that the Consortium cut score should be maintained at **2.57** for level 'G'.

Table 14. Descriptive Summary of Cut Score by Element

	Possible no. of combinations	Mean	Pass		Fail		Total
			n	%	n	%	n
Element 1	335	2.531	1981	*98.85%*	23	*1.15%*	2004
Element 2	66	2.561	1962	*98.40%*	32	*1.60%*	1994
Element 3	45	2.578	1928	*97.03%*	59	*2.97%*	1987
Element 4	16	2.618	2136	*93.11%*	158	*6.89%*	2294
CUT-SCORE		**2.571**					
OVERALL PASS/FAIL DISTRIBUTION (2020)			**1769**	***91.09%***	**173**	***8.91%***	**1942**
OVERALL PASS/FAIL DISTRIBUTION (2019)			***1453***	***91.33%***	***138***	***8.67%***	***1591***

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

4. Moderation and Evaluation

To ensure the fidelity of the assessment and to evaluate the merit, worth and significance of the A*f*GT instrument, a mixed method approach with ongoing validation has been adopted. Consistent with prior years, online cross-institution moderation exercises continued throughout 2020 despite the on-going COVID-19 pandemic and this was further supported by evaluation data collected from PSTs, teacher educators and placement officers from October 2020 to January 2021. These two key activities, the cross-institution moderation and the A*f*GT process evaluation are discussed in sections 4.1 and 4.2 below.

It should be noted that the cross-institution moderation activities are one of two moderation dimensions of the A*f*GT. Prior to the cross-institutional exercise, each institution conducts internal moderation activities within and across their programs of study in accordance with their institution policies to ensure the continuous fidelity and validation of the A*f*GT instrument. Here, the moderation process at cross-institution level is discussed.

4.1 Moderation of A*f*GT

4.1.1 Process

Due to COVID-19, the cross-institution moderation exercises for 2020 were conducted entirely online, following the successful trial of a hybrid model in 2019. The timing of the moderation workshops was also modified from the usual July and December cycle to November 2020 and February 2021. The modification in timing was necessary due to the disruptions faced in placements of PSTs as schools were either closed or had switched to remote or dual learning modalities. As a consequence, a significant number of PSTs' placements were delayed and pushed towards the last quarter of 2020 resulting in the A*f*GT assessment data not being finalised until the end of 2020 and the first quarter of 2021.

Table 15 provides the details of the cross-institution moderation process. The online workshops were designed to collaboratively engage Consortium members in the cross-institution moderation process, while at the same time determining any revisions that may be required for future implementation. In 2020, each institution was represented by at least one lead assessor who participated in both moderation workshops. This process enhancement ensured a more robust process and more consistent data was collected for the cross-institution moderation exercise.

Table 15. Moderation Workshops Details

Moderation Workshop	November 2020	February 2021	TOTAL
Participation*			
Number of assessors	12	12	12
Number of Consortium institutions	10	10	10
Moderated Scripts			
Number of moderated scripts**	36	36	72
Number of blind assessments***	171	111	282

** The assessors and participating institutions were the same for both cycles of the moderation exercise*

*** Moderated scripts denote sample scripts moderated by assessors.*

**** Blind assessments denote assessors marking individual sample scripts. Because each sample script is marked more than once, there are more blind assessments than moderated scripts.*

The following approach was adopted for the cross-institution moderation exercise:

- Institutions were invited to provide scripts for three categories of performance; high, medium and low,
- Scripts were de-identified and randomly assigned to assessors. Each script was either double or triple blind marked,
- Each assessor was assigned a range of scripts identified as high, medium and low performance, and
- During the online workshop, assessors of the same script discussed and moderated the scores and reached a consensus on the final score for the script. The assessors documented their discussion, considerations for changes as well as the agreed final score on a moderation sheet. The objective of this exercise was to determine issues of inter-rater reliability and internal consistency with respect to the instrument.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

4.1.2 Findings on inter-rater reliability

To provide an overview of the moderation data, the descriptive statistics for the two moderation rounds are provided in Table 16. The left-hand table shows data from the first round of moderation (Nov 2020) whilst the right-hand table shows data from the second round of moderation (February 2021).

A total of 72 scripts were moderated, each by at least two assessors whilst most of the scripts were moderated by four assessors. This is indicated by the ***'no. of assessors'*** column. Two scripts (S_06007 and S_07003), were used for norming purposes and hence, they were rated by all the assessors during the online workshop.

The ***'overall score'*** for each script is computed based on the average score awarded by each assessor by element, and then an average of the four elements. The raw grades, 'U', 'G-', 'G', 'G+' were converted to 1, 2, 3, 4, respectively. This means an overall score of 4 would indicate that the PST achieved 'G+' grade on all items in the instrument. The scripts are sorted in descending order from the highest overall score (indicating the highest performing script) to the lowest overall score (indicating the lowest performing script) for each moderation round.

Table 16 also shows the difference between the highest and lowest score for each script grouped by element. A difference of 1 represents one grade level, for example between a 'G+' and a 'G' or between 'G' and 'G-'. A difference of 2 represents a difference of two grade levels, for example a 'G+' and a 'G-'. The data are visually represented using a heatmap, with green indicating little to no difference between the highest and lowest score, yellow indicating some difference and red indicating greater difference. Little to no difference would suggest good inter-rater reliability, whilst large differences suggest poor inter-rater reliability. Due to the variability in the degree of agreement across the scripts for each of the four elements, the moderation data are presented by element rather than on an overall basis.

A couple of observations can be made from the data. Firstly, there is significant improvement in the degree of agreement from the November 2020 round to the February 2021 round. This can be seen both from the heatmap (more green in February 2021 data) and in the average difference for each element. In November 2020, the average difference ranged from 0.81 to 1.1 whilst in February 2021, the average difference decreased to a range of 0.57 to 0.99. This implies that the moderation process facilitated a more consistent approach to marking the A*f*GT scripts amongst the assessors as the exercise progressed from November 2020 to February 2021.

The second observation is that there is better strength of agreement for higher performing scripts relative to low performing scripts. This is consistent with prior years' observations and provides support for the hypothesis that it is easier to agree on a score for high performance submissions, whilst there tends to be more variability for low performance submissions. It is noticeable from Table 16 that as the scripts move from a high overall score to a low overall score, the degree of agreement decreases.

In 2020, it was also possible to perform item analysis to determine internal consistency of the moderation data because the assessors were the same for both cycles of the moderation exercise. The data on internal consistency are presented from two facets, the assessors' view and the item view in sections 4.1.3 and 4.1.4 respectively.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

Table 16. Descriptive Statistics of Moderation Data

Data from November 2020 moderation round

Script	No. of assessors	Overall score	Difference between the most and the least severe ratings			
			E1	E2	E3	E4
S_02007	4	3.67	0.31	0.14	0.67	0.25
S_03013	4	3.65	0.69	1.14	1.00	1.00
S_03007	4	3.56	0.85	0.71	0.83	1.00
S_02005	3	3.44	0.23	0.57	1.17	0.25
S_03002	4	3.43	0.77	1.00	0.67	0.75
S_06006	4	3.39	0.77	0.43	0.33	0.50
S_03011	4	3.34	0.62	1.29	1.50	0.75
S_01012	5	3.24	0.77	1.00	0.83	1.00
S_04002	4	3.16	0.46	0.86	1.33	1.25
S_01009	5	3.15	0.92	1.00	1.50	0.50
S_01011	5	3.08	0.54	1.57	0.83	1.25
S_07002	4	3.07	1.46	0.43	0.67	0.50
S_02002	5	3.07	0.46	0.43	0.50	1.00
S_06005	3	3.04	0.31	1.43	0.67	1.00
S_05003	4	2.96	1.00	0.86	0.67	0.50
S_08001	4	2.93	0.85	1.57	1.17	1.25
S_01013	4	2.92	0.85	0.71	0.67	0.50
S_03001	4	2.87	0.62	1.43	0.67	0.50
S_04001	3	2.86	0.31	0.72	0.50	0.75
S_03012	3	2.85	1.23	1.14	1.17	1.50
S_02001	9	2.83	1.15	1.71	0.50	2.50
S_05001	4	2.74	0.85	0.71	1.17	2.00
S_11003	4	2.67	0.77	1.14	0.67	1.25
S_07004	4	2.66	0.85	1.71	1.67	1.00
S_02004	5	2.64	0.85	1.57	0.50	0.75
S_06001	4	2.63	0.39	1.86	0.50	1.50
S_02006	4	2.63	0.23	1.43	1.17	1.25
S_09003	4	2.61	0.54	0.29	0.83	1.50
S_01008	4	2.61	1.08	2.00	0.50	0.50
S_07001	4	2.58	1.62	1.00	1.33	1.50
S_08003	5	2.57	0.69	1.14	1.50	1.00
S_06007	13	2.47	1.62	1.14	1.17	1.75
S_02003	5	2.45	1.23	1.29	1.17	1.25
S_08002	3	2.44	0.85	1.00	1.33	0.50
S_01010	5	2.40	0.85	1.00	0.67	2.50
S_07003	13	2.06	1.92	2.00	1.50	2.75
Average Difference			**0.82**	**1.10**	**0.93**	**1.10**

Data from February 2021 moderation round

Script	No. of assessors	Overall score	Difference between the most and the least severe ratings			
			E1	E2	E3	E4
S_01019	3	3.69	0.08	0.29	0.83	0.50
S_03010	3	3.46	0.54	0.14	1.00	0.75
S_01004	4	3.44	0.69	0.29	0.33	0.50
S_09005	4	3.43	0.69	0.71	0.33	1.00
S_02053	2	3.36	0.69	0.00	0.00	0.00
S_04010	2	3.30	0.15	0.43	0.17	0.50
S_03006	3	3.29	0.31	0.57	0.33	0.25
S_02036	4	3.29	0.54	0.71	0.83	1.25
S_02037	4	3.27	0.62	0.43	0.67	1.00
S_11001	3	3.15	0.31	1.29	2.00	1.50
S_08008	2	3.04	0.46	1.29	0.83	1.50
S_02052	2	2.94	0.08	0.43	0.50	2.75
S_06002	3	2.84	0.85	0.71	0.83	2.25
S_11002	4	2.82	1.08	1.29	2.50	0.75
S_09002	3	2.82	0.15	1.00	0.50	1.75
S_01016	3	2.81	0.54	0.57	0.33	0.50
S_01017	2	2.81	0.23	0.86	0.17	0.50
S_03018	2	2.78	0.39	0.29	0.33	0.75
S_01005	3	2.73	1.62	1.29	0.17	0.50
S_02008	4	2.72	0.62	1.00	0.67	0.50
S_06008	3	2.70	1.00	0.29	0.67	0.75
S_05007	4	2.58	0.92	1.29	0.83	1.25
S_06003	2	2.56	0.08	0.57	0.00	1.50
S_02018	2	2.47	0.23	0.14	1.17	0.25
S_03017	3	2.45	0.69	1.57	0.67	0.75
S_06004	4	2.43	0.54	0.14	1.67	2.00
S_08004	4	2.41	0.85	1.14	0.17	1.25
S_02035	4	2.40	1.46	0.86	0.83	1.00
S_02019	3	2.38	0.31	0.86	0.50	1.25
S_05004	2	2.38	1.08	0.29	0.83	1.00
S_09004	4	2.28	0.23	1.14	1.00	0.75
S_04003	3	2.26	0.46	0.86	0.50	0.75
S_03016	3	2.13	0.62	1.14	0.83	1.25
S_03014	3	1.87	0.85	0.86	0.17	1.00
S_01015	3	1.81	0.69	0.71	0.83	1.75
S_01014	2	1.80	0.08	0.71	0.33	0.00
Average Difference			**0.57**	**0.73**	**0.68**	**0.99**

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

4.1.3 Internal consistency of moderation data - assessors' view

Figure 15 shows the item response analysis of the 2020 A*f*GT moderation data from the assessors' view. The data are reported by element and by assessor, who were each assigned an individual assessor code. The ***'Measure'*** columns indicate the estimated severity or strictness of the assessor for the element. The higher the measure logit, the stricter the assessor for the element. Each element is sorted from the strictest to the most lenient assessor. Across the four elements, the assessors were fairly consistent in their marking. Relative to their peers, those who were strict were consistently strict and those who were lenient were consistently lenient.

The ***'InfitMS'*** columns show the internal consistency measure for each assessor by element. This goodness of fit statistic measures how close the actual score awarded by the assessor was from the expected score computed using Rasch modelling. Infit values within the range of 0.7 and 1.3 would normally be regarded as 'good fit' while a range of between 0.5 and 1.5 is regarded as acceptable fit. From Figure 15, the internal consistency for all assessors were within the 'good fit' range, with only two instances falling outside of the range. These are highlighted in yellow, for assessor 'A07_2020' in Element 2 and assessor 'A01_2020' in Element 4. Nonetheless, both were still within the 'acceptable' range.

Overall, the assessors who participated in the moderation process showed high internal consistency in their marking, providing further support for the inter-rater reliability of the instrument and the moderation process.

	ELEMENT 1			ELEMENT 2			ELEMENT 3			ELEMENT 4		
	Asssessor	**Measure**	**InfitMS**	**Asssessor**	**Measure**	**InfitMS**	**Asssessor**	**Measure**	**InfitMS**	**Asssessor**	**Measure**	**InfitMS**
STRICTEST	A04_2020	-0.23	0.7	A01_2020	0.61	0.8	A01_2020	0.62	1.2	A01_2020	0.58	1.5
	A01_2020	-0.31	1.3	A04_2020	0.09	0.7	A11_2020	0.35	0.9	A11_2020	-0.29	0.6
	A12_2020	-0.32	0.9	A11_2020	0.08	0.9	A12_2020	0.09	0.8	A12_2020	-0.53	0.8
	A11_2020	-0.42	0.8	A05_2020	-0.28	0.8	A02_2020	0.09	1.0	A05_2020	-0.83	0.6
	A05_2020	-0.53	0.8	A12_2020	-0.32	0.8	A05_2020	0.04	1.0	A04_2020	-0.97	0.7
	A08_2020	-0.60	0.8	A03_2020	-0.37	1.1	A04_2020	0.01	0.9	A07_2020	-1.22	1.3
	A02_2020	-0.62	1.1	A13_2020	-0.44	1.1	A13_2020	-0.20	0.8	A13_2020	-1.38	1.0
	A07_2020	-0.85	1.3	A07_2020	-0.45	1.4	A03_2020	-0.25	1.1	A02_2020	-1.45	1.2
	A06_2020	-0.87	1.0	A02_2020	-0.51	1.1	A10_2020	-0.40	1.2	A06_2020	-1.53	1.1
	A13_2020	-0.96	1.0	A08_2020	-0.71	0.7	A07_2020	-0.49	1.1	A08_2020	-1.64	0.4
MOST	A03_2020	-1.07	1.1	A06_2020	-1.11	1.1	A08_2020	-0.63	1.0	A03_2020	-1.67	0.8
LENIENT	A10_2020	-1.24	1.1	A10_2020	-1.36	1.3	A06_2020	-0.75	0.9	A10_2020	-2.11	1.0

Figure 15. Internal consistency of moderation data - assessors' view

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

4.1.4 Internal consistency of moderation data - item view

Figure 16 shows the item response analysis of the 2020 A*f*GT moderation data from the item view. The data are reported by element and by task, where the ***'Measure'*** columns indicate the estimated difficulty of the task in logit. The higher the measure logit, the more difficult the task. As an example, for the sample of 72 scripts, assessed by the 12 lead assessors, the task with the highest difficulty level in Element 1 is Task 6(c) and the task with the lowest difficulty level is Task 4(e).

The ***'InfitMS'*** columns show the internal consistency measure for each task. This goodness of fit statistic measures how close the actual score awarded for each task was from the expected score computed using Rasch modelling. Infit values within the range of 0.7 and 1.3 would normally be regarded as 'good fit' while a range of between 0.5 and 1.5 is regarded as acceptable fit. As shown in Figure 16, the internal consistency for all tasks were within the 'good fit' range, except for two tasks, Element 1 Task 4(c) and Element 3 Task 1.

The data suggest that the scores for these two tasks are inconsistent with their predicted scores based on the difficulty level of the task and how the assessors have scored the rest of the scripts. This suggests the two items are worth further investigation to understand why their scores are behaving in this unpredictable manner. There may be issues with the task description or the rubric, or the alignment between the task, rubric and standard that are causing inconsistent interpretations among the assessors. Having this moderation data provides an empirical basis to pinpoint where revisions might be required for the A*f*GT. Overall, there was high internal consistency in the task scores except for two tasks, Element 1 Task 4(c) and Element 3 Task 1.

ELEMENT 1		
Item	**Measure**	**InfitMS**
E1_1	-0.36	1.0
E1_2	0.37	1.2
E1_3	-0.54	1.3
E1_4a	0.10	1.0
E1_4b	-0.15	0.8
E1_4c	-0.78	1.7
E1_4d	-0.32	1.0
E1_4e	-0.38	1.0
E1_4f	0.48	1.0
E1_5	0.51	1.0
E1_6a	-0.09	0.7
E1_6b	0.22	0.7
E1_6c	0.92	0.9

ELEMENT 2		
Item	**Measure**	**InfitMS**
E2_1a	0.23	0.9
E2_1b	0.04	0.9
E2_1c	-0.26	1.0
E2_1d	-0.12	0.9
E2_1e	0.24	1.1
E2_2(i)	-0.54	0.9
E2_2(ii)	0.41	1.3

ELEMENT 3		
Item	**Measure**	**InfitMS**
E3_1	0.68	1.7
E3_2	-0.36	0.9
E3_3	-0.53	0.7
E3_4	-0.09	0.9
E3_5	0.05	1.1
E3_6	0.26	0.8

ELEMENT 4		
Item	**Measure**	**InfitMS**
E4_1	0.30	1.0
E4_2	-0.08	1.1
E4_3	-0.12	1.0
E4_4	-0.09	0.9

Figure 15. Internal consistency of moderation data - items view

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

4.2 A*f*GT Process Evaluation

The 2020 process evaluation data were collected through online surveys of PSTs, teacher educators, placement officers and a focus group interview with PSTs. The process evaluation is designed to collect information on the implementation of the instrument and any associated challenges from participants directly involved in the implementation of the A*f*GT. In 2020, these challenges included bushfires and the COVID-19 pandemic and their impact on the implementation of the A*f*GT.

The evaluation survey was distributed within individual institutions to maximise engagement and response rate. At the end of the survey, there was an opportunity for participants to express their interest to participate in a follow-up interview/focus group, which led to one focus group conducted for the PST participant group.

The quantitative survey data were processed and analysed for trends and correlations whilst the focus group transcript and quotes from the survey were thematically analysed and key quotes were extracted to represent the findings. The following sections presents the information collected from the process evaluation data. Due to the small sample size of the placement officers' group, these have been combined with the teacher educator group when reporting the findings.

4.2.1 Teacher Educators and Placement Officers

For teacher educators (which includes academics, course coordinators, lecturers, tutors, and clinical specialists) and placement officers, the survey was designed to explore the respondents' views on:

- Suitability of the A*f*GT as a measurement of readiness to teach,
- Suitability of A*f*GT in a school setting and as an assessment in an ITE program,
- Support provided / received and time commitment during use of the A*f*GT, and
- Operationalising and implementing the A*f*GT as assessment.

8 teacher educators and two placement officers responded to the evaluation survey. The respondents represented the various program types including Masters of Teaching (Secondary), Masters of Teaching (Primary), Bachelor of Education and Graduate Diploma of Education. They also had various involvement in the A*f*GT including preparing PSTs for the A*f*GT, organising suitable schools and other settings for PSTs and assessing the A*f*GT. The sample of responses collected was insufficient to perform quantitative analysis. Instead, a summary or indicative thematic analysis against verbatim material is offered.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

When teacher educators were asked if the A*f*GT adequately measure the planning, teaching, assessing and reflecting aspects of teaching practice, 75% of the respondents said yes. Most teacher educators found the A*f*GT provided a structure and a good yardstick because it exposed PSTs to strategies and skills which assist in addressing different area of teaching and learning. However, there were concerns that the A*f*GT's word limit restricted comprehensive reflections, and did not provide room for creativity and the holistic development of a narrative. Table 17 provides some verbatim quotes from the teacher educator respondents.

Table 17. Feedback on A*f*GT's overall impact on teaching practice

The elements of the A*f*GT require pre-service teachers to critically reflect on the impact of their planning, teaching and assessing on school students' learning. Please make some comments about the focus of the AfGT being on the impact on school students' learning.
It is appropriate that the A*f*GT has a significant impact on student learning as this is a basic fundamental area for classroom practitioners. The A*f*GT exposes students to strategies and skills which assist in addressing this area of teaching and learning.
This is very difficult in [institution] early years environments and somewhat contradicts the [institution] philosophy. Pre-school children can be taught a concept but not evidence that they have understood this until days/months/weeks later. It's the nature of child development. For a PST to assess and evidence that it is 'her' teaching that has impacted a child learning is an issue.
This is vital to ensure that graduate teachers can demonstrate the impact they are having on school student's learning. It also helps graduate teachers to understand the importance of drawing on data to evaluate their practices and the impact of these.
It's a good yardstick for pedagogical adjustments and in the use of assessment but misses the mark when it comes to evaluate and determine connection with students. It actually places students in a passive role.
Encouraging critical reflection is a good thing, and the A*f*GT reflects this to some extent. However links between students learning and the A*f*GT are less clear in some aspects, particularly those where the element title, description of the element and rubric are unclear or do not match.
Given the Clinical Teaching Model is the basis of the [institution]'s Master of Teaching it is absolutely appropriate that the PST's should consider the impact of their teaching (planning & delivery) on student learning.
The A*f*GT is explicit in addressing many elements of the final placement. At times, PSTs misinterpreted the question and did not refer specifically to their placement but instead to the theory/research. The rubric was very explicit in asking PSTs to acknowledge evidence of their students learning. Many PSTs used powerful pedagogical tools but some struggled to subsequently collate evidence of the student learning.

When asked to comment on the contribution the A*f*GT provided for PSTs' professional learning, teacher educators' responses were varied, with some noting that the A*f*GT assessment cannot take the place of the initial teacher education program or critical reflection and that it requires significant support and guidance from teacher educators. Others viewed the A*f*GT as an effective way to emphasize the sequential and cyclical nature of teaching and learning as noted in Table 18.

Table 18. Feedback on A*f*GT's impact on PSTs' professional learning

Please make some comments about the contribution the A*f*GT provided for the participants' professional learning.
The students have just completed a comprehensive 2-year program so I don't think completing the A*f*GT really contributed particularly to what they had already evidenced and learnt over the past 2 years.
They need to be guided to focus more on the outcome for students rather than just their own planning and practice.
It's too long and lacks a sense of equity more contingency to support students is required for instance where students are not permitted to film etc.
Many PST's come to see the A*f*GT as a 'check box' activity where they need to write 50 words which address the highest level on the rubric. While they may learn something from this, it may devalue the process of engaging in critical reflection.
The A*f*GT emphasises the sequential nature and cycle of teaching and learning. It also really shows how the teaching is linked to student learning - a concept that is not always picked up by new teachers who can be heavily focused on their lesson delivery. Student learning can be lost in the intensity of lesson planning and curriculum coverage.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

Turning to support provided to teacher educators in implementing the A*f*GT, teacher educators found the support positive. Figure 17 shows that 70% of the teacher educators rated the support received were extremely good, and 80% felt that there was extremely good ongoing support provided for the A*f*GT implementation.

When asked about issues encountered for A*f*GT, teacher educators cited not having exemplars and samples of work available for PSTs as problematic, and created stress for the PSTs.

> *Element 4 is a "stand alone" Element which is not part of the Workbook. This is problematic for a number of reasons. There are no exemplars/samples of completed/partially completed Workbooks available to students. This causes a great deal of stress for students who rely upon university staff for guidance and support regarding the format and structure of a completed Workbook.*

Some teacher educators also cited better alignment and clarity could be provided in the *AfGT Information Guide*:

> *Element 1 part 4 could be more concise. For example, an overall statement to justify teaching resources rather than repeating this for each lesson. The task instructions and rubric could be better aligned (if the instruction is to evaluate then G on the rubric should be evaluate).*
>
> *The clarity around research (eg. E1 '3') meant that PSTs sometimes misinterpreted this and discussed research without specific contextual information.*

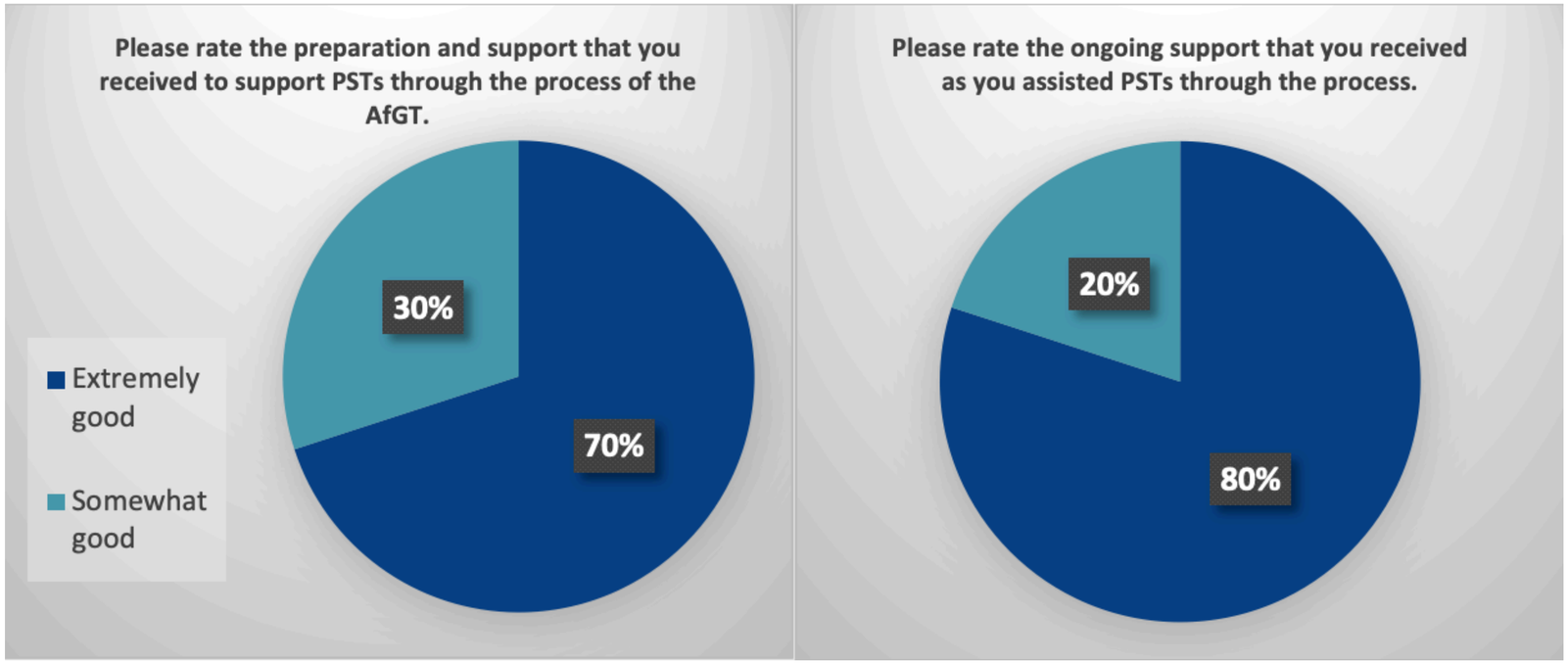


Figure 17. Support received by teacher educators and placement officers

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

Significantly, in 2020, most teacher educators cited COVID-19 as a major challenge that impacted the implementation of the A*f*GT, including the difficulty in sourcing placement, school closure, online and blended delivery and shortened timelines.

> *This had adverse effects on a number of PSTs. It also tested the pedagogical expertise and adaptability of teachers who had to teach remotely for the first time. Some were absolutely outstanding - this may also have been indicative of the extra support they received from their mentor teacher. As a teacher educator, I took it upon myself to add in some demonstrations of teaching via zoom and using an LMS in a secondary school.*
>
> *Many placements cancelled, delayed and shortened, which potentially had an effect on AfGT outcomes. All PSTs worked hard to ensure their work was not compromised by the issues they faced.*

Overall, despite the challenges faced in 2020 due to the COVID-19 pandemic, teacher educators and placement officers regarded the A*f*GT as a valid and coherent teaching performance assessment instrument. It is notable that the feedback is now more precise around specific implementation challenges and on specific areas in the instrument which require attention. The level of insight and familiarity with the requirements of A*f*GT expressed by teacher educators provides support that the A*f*GT is an established, mature assessment that is subject to continuous review and evaluation. Opportunities for more advanced resources such as annotated examples and capacity building infrastructure such as assessors' training and may be considered by the Consortium in the future.

Photo by Kelly Sikkema on Unsplash

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

4.2.2 Pre-Service Teachers (PSTs)

The survey for PSTs explored more detailed aspects of the A*f*GT from a user perspective including the clarity, appropriateness and difficulty of each A*f*GT element as well as PSTs' feedback on the guidance materials provided. Table 19 provides a descriptive summary of the PSTs who completed the evaluation survey. A total of 87 PSTs from five institutions responded to the survey in 2020. Both undergraduate (46 PSTs) and postgraduate (41 PSTs) degrees were represented. The number of professional experience placement days undertaken by the respondents ranged between 11 and 60 days with Figure 18 showing the distribution of placement days.

Table 19. Survey Responses from PSTs

Institution	No. of complete responses
University 1	51
University 2	21
University 3	9
University 4	4
University 5	2
Total	**87**
Program Type	**No. of complete responses**
Bachelor of Education	8
Bachelor of Education (Early Childhood)	9
Bachelor of Education (Primary)	22
Bachelor of Education (Secondary)	5
Graduate Diploma of Education	2
Masters of Teaching	6
Masters of Teaching (Early Childhood &	1
Masters of Teaching (Early Childhood)	2
Masters of Teaching (Primary)	6
Masters of Teaching (Secondary)	26
Total	87

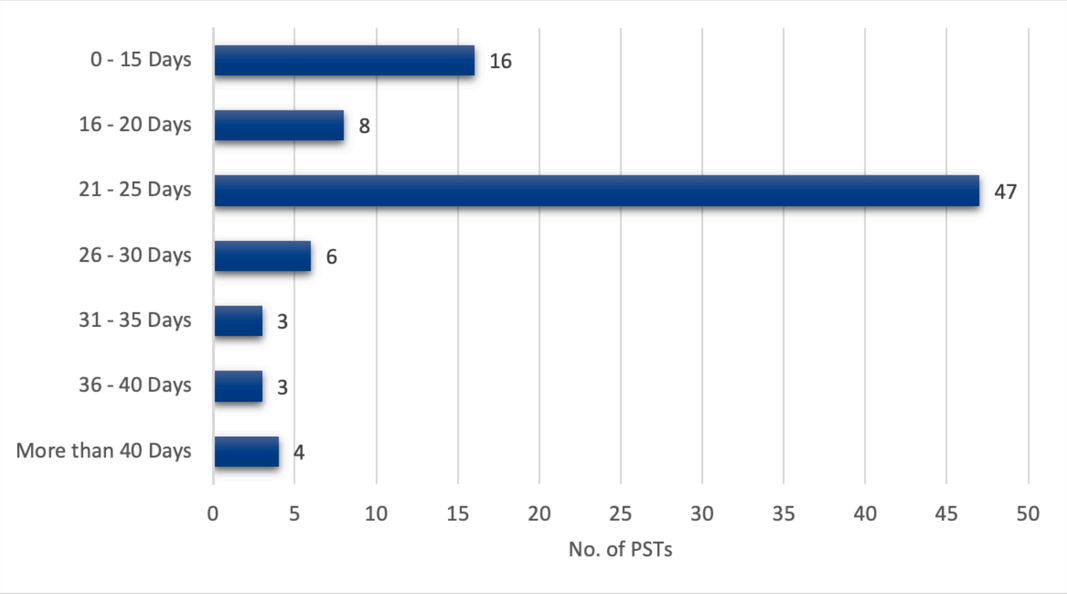


Figure 18. Distribution of placement days

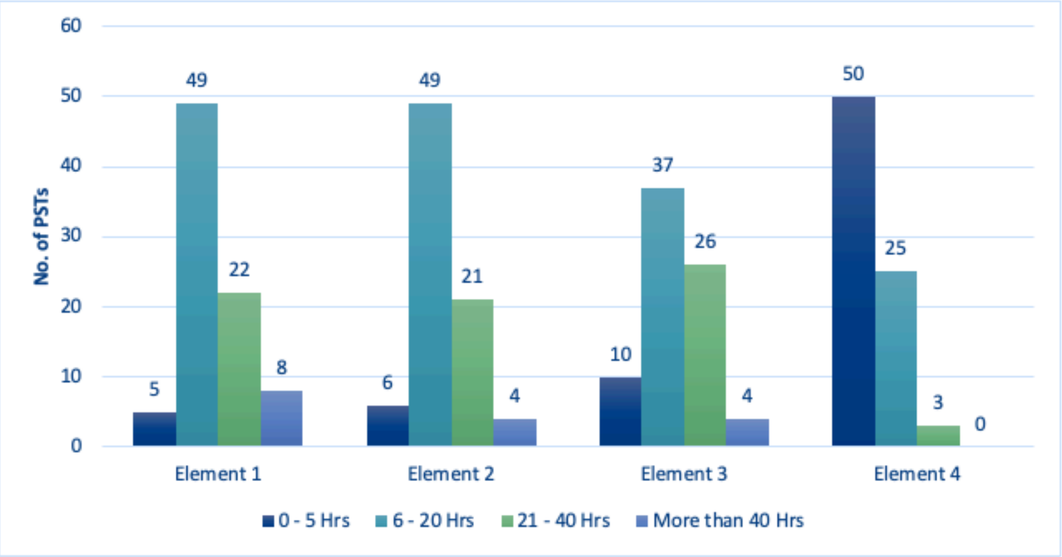


Figure 19. Time taken to complete each element

When asked about the time spent (in hours) to complete each A*f*GT element, Figure 19 shows most PSTs indicated that they spent between 6 and 20 hours for Elements 1, 2 and 3 and between 0 and 5 hours for Element 4. As there is a 24-hour limit for PSTs to complete the Element 4 online assessment, it is no surprise that this element took the least time to complete. For Elements 1 to Element 3, the time taken by PSTs to complete each element was fairly consistent, with Element 3 reportedly taking slightly more time than Elements 1 and 2.

The following figures represent PSTs' response to the four A*f*GT elements in terms of clarity (Figure 20), relevance (Figure 21) and degree of difficulty (Figure 22).

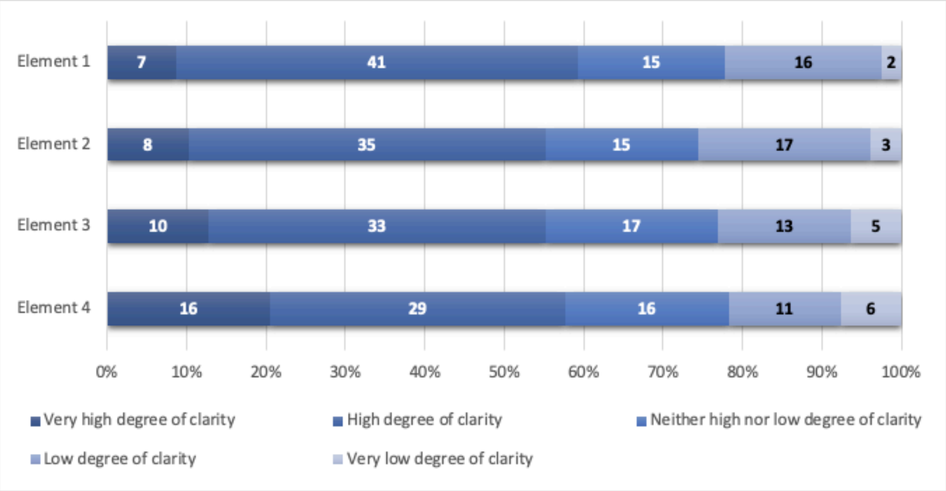


Figure 20. Clarity of task

57% of PSTs responded favourably (high to very high) to the survey question on the clarity of the task (Figure 20). This is a significant increase from the 2019 PSTs' survey responses, where only 37% responded favourably. This seems to suggest that a majority of the PSTs felt they were adequately supported in accessing the requirement of the A*f*GT. Between each element, PSTs found Element 2 having the least clarity (26% negative response), and Element 4 having the most clarity (22% negative response).

MacIver et. al, (2014) assert that clarity of task does not equate with what is termed as 'user validity'. To appreciate user validity, questions were asked concerning the relevance of the four A*f*GT elements (Figure 21). Consistent with 2019 responses, most PSTs (71%) responded favourably when asked about the relevance of the A*f*GT tasks (2019: 66% PSTs). Between each element, Element 2 had the most PSTs responding as irrelevant (23% negative response), and Element 3 was deemed by PSTs to be the most relevant element to teaching (13% negative response).

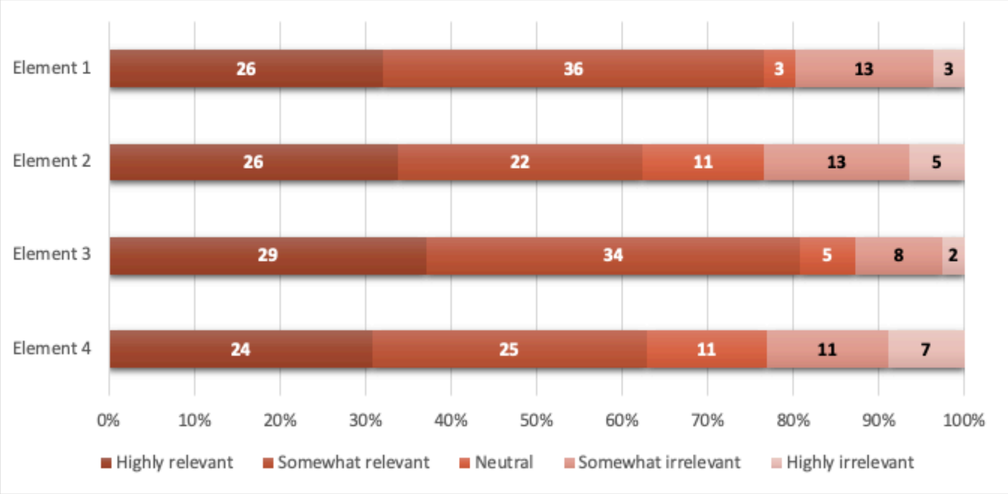


Figure 21. Relevance of task

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

PSTs were also asked to rate the degree of difficulty for each A*f*GT element (Figure 22). 40% of the PSTs found the A*f*GT to be difficult (high and very high degree of difficulty) and 39% provided a neutral response. This is again consistent with 2019 survey responses where 45% of the PSTs found the A*f*GT to be difficult and 44% provided a neutral response. Between each element, Element 4 had the most responses for low degree of difficulty (24%) whilst Element 3 was found to be the most challenging (15% low difficulty response). In 2019, Element 4 also had the most responses for low degree of difficulty (15%), but PSTs found Element 2 as the most challenging element (7% low difficulty response). Overall, a greater proportion of PSTs were finding the A*f*GT less difficult when compared with 2019 data.

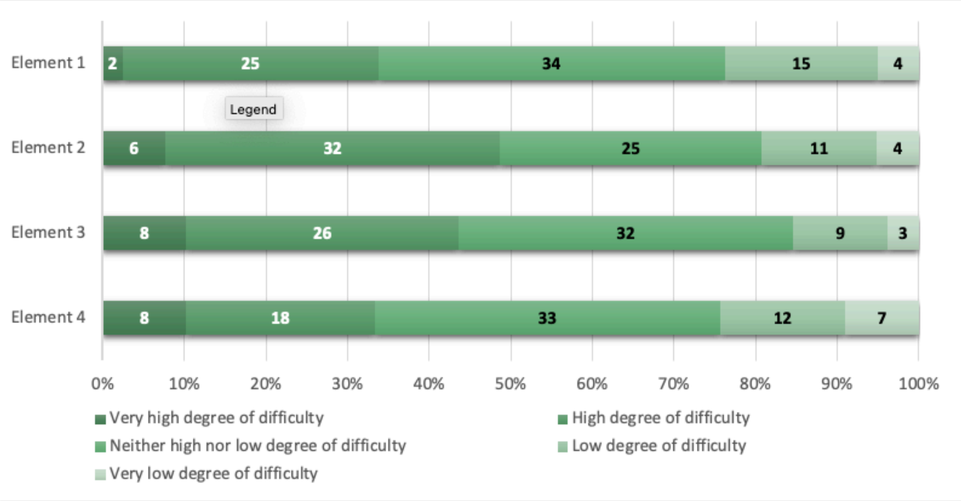


Figure 22. Degree of Difficulty

In 2020, PSTs were asked if they used the *AfGT Information Guide* and the *AfGT Manual* when preparing their responses for the A*f*GT assessment tasks. This question is asked to provide context for the next set of questions regarding the clarity of materials provided to support PSTs. As expected, an overwhelming majority of the PSTs used the *AfGT Manual* and the *AfGT Information Guide*. However, a greater proportion of PSTs referred to the *AfGT Manual* than the *AfGT Information Guide* as reflected in Figure 23.

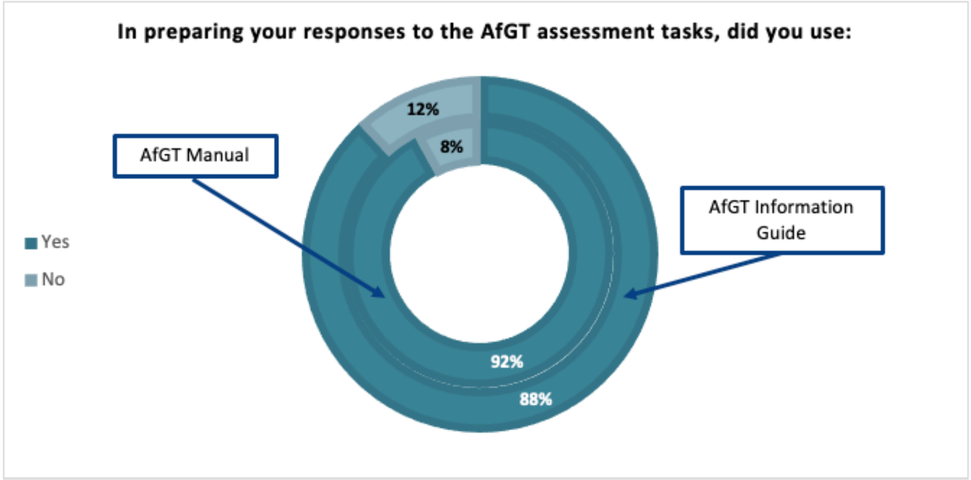


Figure 23. Usage of A*f*GT materials

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

When asked about their perception on the guidance materials provided for the A*f*GT and the extent to which PSTs found the A*f*GT a coherent assessment of their teaching practice (Figure 24), most PSTs responded favourably. Namely, 52% said there was a high to very high level of clarity for guidance materials and 55% perceived the A*f*GT to be highly to very highly coherent. Only 5% of the PSTs felt there was very low level of clarity in guidance materials and level of coherence in the A*f*GT assessment.

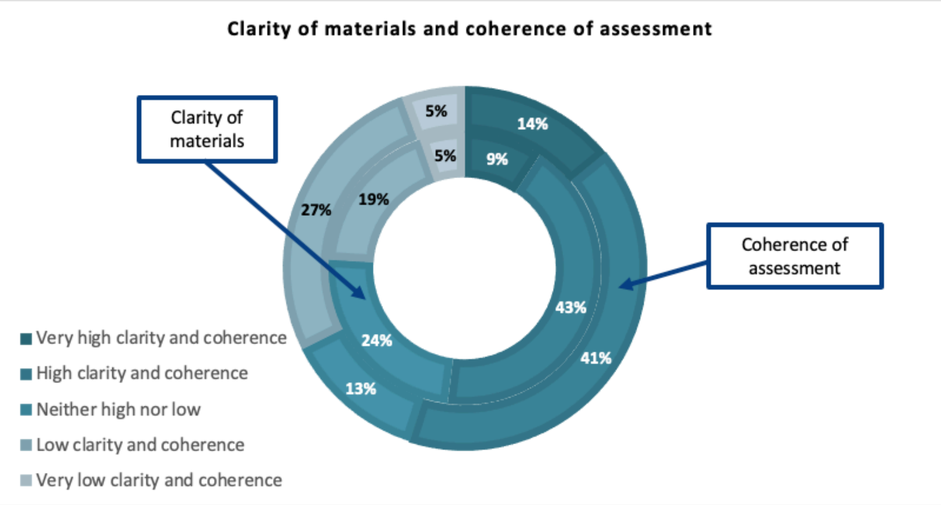


Figure 24. Clarity of materials and coherence of assessment

For all the dimensions surveyed, the 2020 PST data were more favourable compared to previous years. In 2020, more PSTs agreed that the A*f*GT was clear, relevant, coherent and had a more manageable degree of difficulty in the tasks. When analysed together, PSTs found the A*f*GT tasks to be both coherent and challenging. They also perceived the A*f*GT assessment as relevant and as an appropriate indicator of their classroom readiness. Overall, this provides support for the merit of the instrument, and reflected the complex and challenging intellectual work of teaching. It is also worth noting that these self-assessed responses should be viewed alongside the findings from the actual scores or grades obtained by PSTs (see Section 3) to provide a fuller picture, as a perception that an assessment is difficult does not necessarily mean a poor result.

4.2.3 Impact of COVID-19 on PSTs

In 2020, we took the opportunity to gather data on unforeseen events that hindered or interrupted PSTs' completion of the A*f*GT. 77% of the PSTs found COVID-19 to be a major challenge, 1% identified the bushfires (from late December 2019 to early 2020) and 5% identified other circumstances such as personal and family issues, health issues and lack of final placement schools. Interestingly, 17% of the PSTs did not find any circumstances or events that hindered the completion of the A*f*GT.

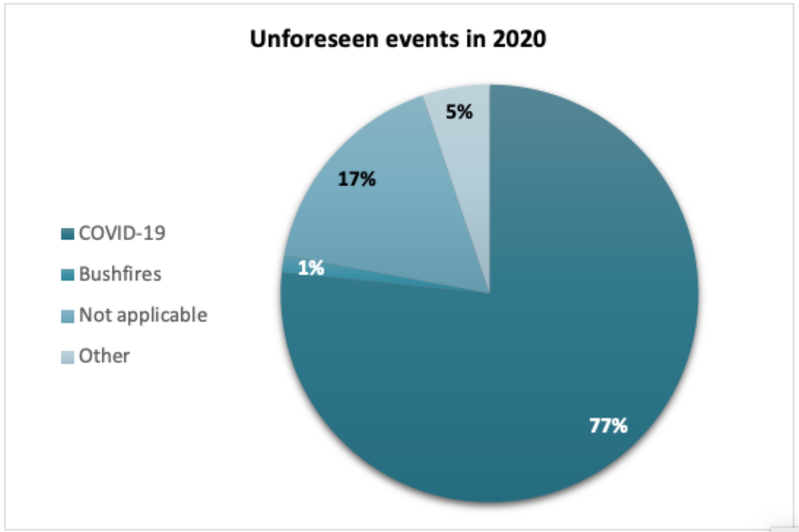


Figure 25. Unforeseen events identified by PSTs that impacted their completion of the A*f*GT

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

When asked to comment about the unforeseen events, PSTs expressed the challenge of having a shortened placement and remote learning as the major hindrances to the completion of the A*f*GT:

> *Covid 19 changed the situation within the classroom. It changed many procedures and the way things happened in the classroom. It also changed the length of my placement - this meant we had a LOT of work to do in a smaller space of time.*
>
> *My practical placement was reduced by 5 weeks. Because I have worked in education as an Education Assistant for 15 years, this did not impact my readiness as much as it would have impacted a younger school leaver graduate teacher.*
>
> *Completing the AfGT over a shorter placement really put me under the pump as the placement got more intensive straight away.*
>
> *Lockdowns and online learning had a major impact on fulfilling the requirements in time to graduate.*
>
> *Covid only to the extent that the placement was reduced to 5 weeks instead of 10 and therefore the task of completing the AfGT in a shorter period took the emphasis away from enjoying the placement and more on completing the AfGT as there was limited time. The assessment may should have been reduced somehow and maybe something that can be done in the future.*
>
> *The shortened length of the final placement due to COVID-19 did make the final practicum quite intense and fast-paced.*
>
> *The completion of AfGT, [another subject] and research project in the final semester, as a result of COVID, negatively impact my overall enjoyment and performance. The scheduling of these time-consuming tasks in an overlapping window increased my overall stress levels and negatively impacted my ability to complete the tasks to a degree of my usual satisfaction. Further, this also impacted my ability to focus on completion of key selection criteria and future job employment.*

Whilst the A*f*GT continued to be implemented with fidelity throughout the challenging COVID-19 pandemic in 2020, some PSTs had to adapt to the shorter placement period. As expressed in the comments above, this had a negative impact on their professional experience and was an added challenge for the PSTs in completing the A*f*GT.

Photo by Gautam Arora on Unsplash

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

4.3 Summary and Future Considerations for the Instrument

The results and processes described in Section 3 and Section 4 all contributed to the validation of the A*f*GT instrument. Based on 2020 data, the analyses continue to substantiate the A*f*GT as a valid, reliable and fair teaching performance assessment instrument. As outlined in Program Standard 1.2 (AITSL, 2019), the A*f*GT demonstrates the following features:

1.Valid reflection of classroom teaching practice (including planning, teaching, reflecting and assessing student learning):

a.The results reveal that the A*f*GT is a valid reflection of classroom teaching and that the majority of the A*f*GT items are correctly ordered.

b.Collectively, the results demonstrate high consistency in the distribution of grades across the four elements, suggesting consistency in scoring across the elements.

c.The data suggest that there is an opportunity for success in each of the elements and a relatively equivalent score across the elements. This is significant, and provides support that the instrument is robust and highly stable, given the varying number of tasks in each element.

d.The A*f*GT is not showing any systematic bias for the various sub-groups of program type (bachelor, masters, primary, secondary, early childhood, etc.), although the data reveals that PSTs from primary and secondary program type score slightly differently to each other on a small number of tasks in Elements 1 and 3.

e.When teacher educators were asked if the A*f*GT adequately measures the planning, teaching, assessing and reflecting aspects of teaching practice, 75% of the respondents said yes.

2.Valid assessment that assesses the content of the Graduate Teacher Standards:

a.Given the objective of the A*f*GT is to assess PSTs' attainment of the specified APSTs at the Graduate level rather than used as a ranked assessment, the results reveal that the conceptual design of the A*f*GT is a valid assessment of the content of the Graduate Teacher Standards.

b.To pass the A*f*GT, PSTs are required to pass all four elements. While the A*f*GT assesses the content of all the Graduate Teacher Standards, it is possible to identify the items that prove to be most and least challenging to achieve a 'G' or 'G+'.

c.When PSTs were asked how relevant were the A*f*GT task in reflecting the Graduate Teacher Standards, 71% responded favourably.

3.Measurable and justifiable achievement criteria that discriminate between meeting and not meeting the Graduate Teacher Standards:

a.For all four elements, the A*f*GT is highly effective at obtaining precise estimates of PSTs who are 'on-the-cusp', where it is critical to determine if a PST has indeed met the APSTs at Graduate level. Element 4 reflects this particularly well.

b.The A*f*GT items are effective in separating PSTs at the low end of the classroom readiness scale.

c.As part of the ongoing validation process, the cut score was confirmed as representative of the score distribution based on 2020 sample data.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

4.Reliability of scoring between assessors:

a.The distribution of grades for each element across each institution is consistent, with some within-institution variations identifiable.

b.Overall, all the assessors who participated in the cross-institution moderation process showed high internal consistency in their marking.

c.There is better strength of agreement for higher performing scripts relative to low performing scripts, with more variability for low performance submissions.

5.Moderation processes that support consistent decision making against achievement criteria:

a.Consistent with prior years, the inter-rater reliability analysis showed strong consensus among the assessors who participated in the standard-setting activity. Importantly, the assessors achieved stronger levels of agreement as the moderation rounds progressed through the online cross-institution moderation workshops.

b.There is strong evidence to suggest that assessors agree what classroom readiness looks like and performance standard that meets the APST at Graduate level.

c.The current cross-institution moderation process ensures high-quality data are gathered as evidence of validity and reliability of the instrument whilst informing the Consortium on specific areas in the instrument which may benefit from refinement. This ensures the task descriptions remains clear and are contextually responsive as part of the continuous improvement process.

d.Maintaining vigilance on the cross-institution moderation processes will remain a high priority for the Consortium.

To sustain its progress and ensure that the A*f*GT maintains its fidelity, especially in an uncertain and potentially disruptive context, the following areas have been identified as key focus areas for the A*f*GT assessment in the next three to five years.

4.3.1 Continuous Validation Process

The ongoing validation process remains the bedrock of ensuring that the A*f*GT continues to assess classroom readiness as intended. For A*f*GT, validation can be viewed from both an internal and an external dimension.

Internally, the framework for establishing A*f*GT validity and reliability adopted during the design and development phase continues to guide the validation process. The mixed method framework, shown in Figure 26, is a systematic approach to collecting evidence to support the A*f*GT's value and worth. Overall, the A*f*GT has established strong validity, reliability and fairness measures. While many of the methods described in the framework in Figure 26 are quantitative, the exercise of professional judgement is equally critical in determining the validity and reliability of the instrument. Thus, gathering process evaluation data from participants and ongoing discussions among Consortium members within the various Committees will remain a critical part of the validation process.

Externally, the A*f*GT Consortium will continue to engage with AITSL, other TPAs and TPA consortia in collaborative initiatives such as described in Section 2.7 to ensure the AfGT is assessing PSTs' competence against the APSTs consistently. Apart from ongoing moderation and cross-TPA collaborations, continuous discussions with various external stakeholders such as statutory bodies, departments of education and international experts are also important to collectively exercise that judgement.

4.3.2 Continuous Feedback and Refinement Process

One of the key objectives of the A*f*GT's process evaluation is to collect evidence from a broad range of stakeholders to determine if any refinement to the instrument or enhancement to the implementation process is required. This is an iterative and collaborative process of continuous feedback, implementation, evaluation and refinement. Some of the planned activities such as support materials arising from the feedback process are described in the next Section.

4.3.3 Continuous Capacity Building and Quality Assurance Process

To further consolidate inter-rater reliability and ensure consistency of judgement across the Consortium, the third area of focus is building capacity and enhancing the measurement quality of the A*f*GT instrument. This strategy will provide a consistent approach to assessment and moderation across all the institutions, at every level (program, courses and subject).

This is particularly important as new member institutions join the Consortium, or when there are changes in personnel within institutions. To support this, resources and training such as assessors' training and handbook, moderation manuals and 'on-boarding' for new assessors and new institutions will be useful for the Consortium to consider.

Elements	Respondents	Validity evidence				Fairness evidence	Reliability evidence
		Content	Internal Structure	Other variables	Consequential		
1	Pre-service teachers	Review and rating by content experts	Correlations	Other elements of A*f*GT	Process evaluation feedback from participants	Analyse specific groups using mean scores and distribution	Internal consistency
			Factor analysis				
			IRT Analysis	Course grades and mentor teacher ratings	Costs, unintended consequences	Differential item functioning	Inter-rater reliability
2	Pre-service teachers, mentor teachers, peers and students	Review and rating by content experts	Correlations	Other elements of A*f*GT	Process evaluation feedback from participants	Analyse specific groups using mean scores and distribution	Internal consistency
			Factor analysis				
			IRT Analysis	Course grades and mentor teacher ratings	Costs, unintended consequences	Differential item functioning	Inter-rater reliability
3	Pre-service teachers	Review and rating by content experts	Correlations	Other elements of A*f*GT	Process evaluation feedback from participants	Analyse specific groups using mean scores and distribution	Internal consistency
			Factor analysis				
			IRT Analysis	Course grades and mentor teacher ratings	Costs, unintended consequences	Differential item functioning	Inter-rater reliability
4	Pre-service teachers	Review and rating by content experts	Correlations	Other elements of A*f*GT	Process evaluation feedback from participants	Analyse specific groups using mean scores and distribution	Internal consistency
			Factor analysis				
			IRT Analysis	Course grades and mentor teacher ratings	Costs, unintended consequences	Differential item functioning	Inter-rater reliability

Figure 26. Framework for establishing A*f*GT's assessment validity and reliability

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

5. Refinements to the Instrument

Based on the A*f*GT moderation workshops and evaluations conducted at the end of 2020, several minor refinements to the A*f*GT instrument were made to provide clarity and to ensure greater consistency in implementation. A summary of the main changes appears in Table 20. These documents were updated in v1.5:

- *AfGT Information Guide*
- *AfGT Manual*, and
- *AfGT Additional Information for Institutions*.

Continuing from previous years, the Implementation & Improvement Committee (IIC) expanded the Element 4 item bank with a view of eventually retiring – or resting – certain scenarios. In 2020, new scenarios were added, bringing a total of eight scenarios per set. When PSTs undertake Element 4, they are randomly assigned one of the possible eight scenarios, limiting the likelihood of PSTs being presented with the same scenario, and thus reducing the likelihood of plagiarism.

Having a sufficiently large item bank provides the Consortium flexibility to select scenarios that best reflect the standard being assessed using evidence-based data. This is informed, in part, by the statistical analysis conducted for Element 4 as reported in Section 3.3.2 above. Enhancements to Element 4 is anticipated to continue, although the focus for 2021 would be to review and refine existing scenarios rather than developing new ones now that the item bank is established.

Table 20. Summary of Refinements to the A*f*GT Documentation

Issues Identified during Moderation & Process Evaluation	Refinements Made
Element 1: Clarity on the number of lessons in the learning sequence	Learning sequence clarified as no fewer than five lessons and no more than eight lessons.
Element 1: Documenting the impact of COVID-19	Guideline is provided in Element 1 Table 1 where PSTs are encouraged to record – in as much detail as possible – specific instances that have had an impact on their ability to administer the A*f*GT as intended. In the instance of the impact of COVID-19, for example, the PST should make note of the specific things that have prevented them from being able to video record students during periods of remote and online teaching.
Element 1 Task 4: Format not conducive for assessors to review as it requires significant scrolling back and forth	The lay-out of the lesson plans (Element 1 Task 4) was refined so that it is arranged by task (E1-4b, E1-4c, etc) followed by lessons, instead of lessons followed by tasks in the previous version.
Element 2: Video recording – procedures and guidelines that align with A*f*GT's Privacy Policy	Further clarity in the guideline given on video recording, that the attention is on the PST teaching rather than what the students are doing. PSTS are strongly encouraged to use schools' recording equipment and they must avoid capturing information that might identify students.
A*f*GT Privacy Policy	Incorporated into the A*f*GT Information Guide
Updates in wording and terminology	Change references from 'university' to 'institutions', removing the second person pronoun (you, you, etc) in the instructions and ensuring consistency between instructions and rubrics.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

6. Consortium Initiatives

A number of Consortium initiatives are at various stages of development, and all involve the use of technology, as described below.

6.1 The Use of Computers to Support Assessors

At the December 2019 Moderation and Evaluation Workshop, an idea was proposed by the University of Sydney (UoS) about the possibility of the Consortium trialling the introduction of Artificial Intelligence (AI) to assist assessors in making more accurate decisions—in much the same way as assessors utilise text-matching software to assist the judgements they make. Element 4 appeared to provide an ideal starting platform as it was already in a digital format. In order to 'train' the system, it was considered that the trial would need a certain number of completed submitted responses to the scenarios. Upon further examination of the testing methodology, UoS reports that it had enough submissions from their own PSTs to look at the feasibility and accuracy of the research.

The initial results indicate that a Machine Learning model can, to an acceptable degree of accuracy, appropriately determine scores for free text responses utilising the single marking rubric for Element 4 of the A*f*GT. The researchers maintain the hypothesis that with additional test data combined with further investigation of Machine Learning techniques, that the performance of the data model would continue to improve. This initiative provides an opportunity for the Consortium to understand the potential power and practicalities of moving to an online platform.

6.2 Resources to Support Schools and Mentor Teachers

The A*f*GT Consortium is currently in the midst of developing consistent documentation, communication packs and resources to support schools and mentor teachers in the implementation of the A*f*GT. These resources could include short videos which will be more accessible to teachers. The current guidance materials, whilst relevant, requires heavy reading. For time-poor teachers, short video clips that provide brief overviews would be helpful. These video clips will then point them to the guidance materials where they can make further references.

6.3 Providing Institutions with Assessment Feedback

As the number of PSTs undertaking the A*f*GT increases, there are more data upon which to make analytical judgements—one of which relates to the degree of difficulty of marking by assessors in individual institutions. It is therefore possible to provide institutions with customised confidential feedback about how the assessors—as a group, not individually—in their institution are marking in comparison to other institutions within the Consortium. This information will only ever be available to each individual institution; the results are not to be shared in reports to the Consortium.

6.4 Moving Ethics Documentation Online

The A*f*GT Consortium is considering trialling an approach similar to the A*f*GT process evaluation survey where external parties can access an online survey link (Qualtrics) to provide consent to participate in the research component of the A*f*GT. Given that this process works smoothly for process evaluation, it could be feasible to collate research consent forms using an online mode. This could be trialled with PSTs and university-based personnel in the first instance, before expanding to other school-based participant groups such as principals, mentor teacher, parents and students. Consortium members will be invited to participate in discussions in relation to this initiative.

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

7. References

Acar, T. (2011). Sample Size in Differential Item Functioning: An Application of Hierarchical Linear Modeling. *Educational Sciences: Theory and Practice, 11*(1), 284–288.

Australian Institute for Teaching and School Leadership. (2019). *Accreditation of initial teacher education programs in Australia*. AITSL Retrieved from https://www.aitsl.edu.au/tools-resources/resource/accreditation-of-initial-teacher-education-programs-in-australia---standards-and-procedures

BBC News. (2020, October 26). *COVID in Australia: Melbourne to exit 112-day lockdown*. BBC News. https://www.bbc.com/news/world-australia-54686812

Clinton, J. (2020, April 10). *COVID-19 Message from the Chair*. AfGT Consortium LMS. https://canvas.lms.unimelb.edu.au/courses/88880/discussion_topics/186690

Keamy, R. K., & Selkrig, M. A. (2021). Interrupting practice traditions: Using readers' theatre to show the impact of a nationally mandated assessment task on initial teacher educators' work. *Teaching Education*. https://doi.org/10.1080/10476210.2021.1951198

Kriewaldt, J., Walker, R., Morey, V. Morrison, C. (2021) Activating and reinforcing graduates' capabilities: Early lessons learned from a Teaching Performance Assessment. *Australian Education Researcher*. https://doi.org/10.1007/s13384-020-00418-4

Le, L. (2006). Analysis of Differential Item Functioning, Australian Council for Educational Research. https://www.acer.org/files/analysis_of_dif.pdf

MacIver, R., Anderson, N., Costa, A.-C., & Evers, A. (2014). Validity of Interpretation: A user validity perspective beyond the test score. *International Journal of Selection and Assessment*, (2), 149. https://doi-org.ezp.lib.unimelb.edu.au/10.1111/ijsa.12065

McGraw, A., Keamy, R. K., Kriewaldt, J., Brandenburg, R., Walker, R., & Crane, N. (2021). Collaboratively designing a national, mandated teaching performance assessment in a multi-university consortium: Leadership, dispositions and tensions. *Australian Journal of Teacher Education*. http://dx.doi.org/10.14221/ajte.2021v46n5.3

Muraki E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2):159–76. doi:10.1177/014662169201600206

Nguyen, T. H., Han, H. R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *The patient, 7*(1), 23–35. https://doi.org/10.1007/s40271-013-0041-0

Walker, C. M. (2011). What's the DIF? Why Differential Item Functioning Analyses Are an Important Part of Instrument Development and Validation. *Journal of Psychoeducational Assessment, 29*(4), 364–376.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, N.J.: Lawrence Erlbaum Associates, 2005.; cat00006a

Zanon, C., Hutz, C. S., Yoo, H. (Henry), & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica, 29*(1), 18. https://doi.org/10.1186/s41155-016-0040-x

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References

Title Page
Table of Contents
Executive Summary
Introduction
Consortium Update
Findings from 2020 Data
Moderation and Evaluation
Instrument Refinement
Consortium Initiatives
References