Receptivity to Instructional Feedback: A Validation Study in the Secondary School Context in Singapore

Anastasiya A Lipnevich

Queens College and the Graduate Center, City University of New York, USA

Carolina Lopera-Oquendo

The Graduate Center, City University of New York, USA

Citation: Lipnevich, A. A. & Lopera-Oquendo, C. (2022). Receptivity to Instructional Feedback: A Validation Study in the Secondary School Context in Singapore. *European Journal of Psychological Assessment*.

Correspondence concerning this article should be addressed to Anastasiya A. Lipnevich, Department of Educational Psychology, the Graduate Center, The City University of New York, 365 5th Avenue, New York, NY, 10016 USA; e-mail: alipnevich@gc.cuny.edu.

Anastasiya A. Lipnevich https://orcid.org/0000-0003-0190-8689
Carolina Lopera-Oquendo https://orcid.org/0000-0002-9355-5843

This study was funded by Singapore Ministry of Education (MOE) under the Education Research Funding Programme (OER 02/19 KTHK) and administered by National Institute of Education, Nanyang Technological University, Singapore. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Singapore MOE and NIE.

Abstract

The purpose of this study was to report validity evidence for the instrument intended to measure receptivity to instructional feedback in a sample of secondary school students from Singapore (*N* = 314). We tested a nested hierarchy of hyphotesis for addressing the cross-group (i.e., gender) invariance and compared means on the receptivity to feedback sub-scales between gender groups. We also examined whether receptivity to feedback predicted student grades. The four-factor hypothesized model comprising experiential attitudes, instrumental attitudes, cognitive engagement, and behavioural engagement with feedback had a good model fit. Multi-group confirmatory factor analysis supported configural, metric, partial scalar, partial strict as well as variance and covariance invariance across gender groups. After controlling for gender, cognitive engagement and experiential attitudes predicted increments in grades, suggesting evidence for discriminant validity among the receptivity factors as well as their relevance for prediction of meaningful educational outcomes.

Key words: feedback, receptivity, engagement with feedback, validity evidence, secondary school

Receptivity to Instructional Feedback: A Validation Study in the Secondary School Context in Singapore

Introduction

For decades, the main focus of research in the filed of instructional feedback was on the *provision* of information to the student about his or her progress. Specific characteristics of the feedback message were considered to be the *sine qua non* of helping students in their performance and learning. Gradually, the field has shifted to realize that providing feedback only begins the process that leads to student improvement (Jonsson, 2013; Winstone et al., 2017). It is now uniformly accepted by researchers and practitioners alike that feedback is only effective if it is used, and it will only be used if a student has the right context, information, dispositions, and approach to how to use it. Lipnevich et al. (2016) described a model of student/feedback interaction, where they discussed feedback information, student as the recipient of feedback, and how the two interact. One of the key learner characteristics that the authors discussed was student receptivity to instructional feedback.

The idea of student receptivity is based on the premise that there may be individual differneces in the way people are willing or ready (or not) to accept feedback (Lipnevich et al., 2016; Lipnevich & Smith, 2009; Murano et al., 2021). Some students may generally be eager to receive external comments on their progress or performance, whereas others may be less welcoming of it. These differences may be situational and context-dependent, but a general, trait-like feedback receptivity appears to exist. To test this claim, Lipnevich at al. (2021) constructed an instrument to examine student receptivity to instructional feedback. The researchers explored links to the Big Five personality factors of conscientiouseness, agreeableness, neuroticism, openness, and extraversion and showed a general trait component to the receptivity construct

(Lipnevich et al., 2021). That is, receptivity incremented over broader personality characteristics in a sample of university students from the US and New Zealand. Other researchers have found a more situational (state) receptivity disposition among students (Brown et al., 2014) and described related concepts, such as proactive recipience (Winstone et al., 2017).

Lipnevich et al. (2021) described receptivity as comprising cognitive (do I understand feedback?) and behavioral (do I know what to do about it?) engagement, as well as instrumental (do I think it's useful?) and experiential (do I like it?) attitudes. The authors revealed that conscientiousness and openness were the strongest predictors of receptivity, suggesting that students who were achievement-oriented and disciplined as well as intellectually curious and open to new information would tend to be more receptive to feedback. Not surprisingly, neuroticism negatively predicted behavioural engagement with feedback. Agreeableness had weak links with receptivity, and extraversion was not related to feedback receptivity at all. These patterns of relationship was reassuring and suggestive of the need to further explore this construct and its generalizability to different cultures and ages, as well as potential links to student performance indicators.

There is an extensive body of research that describes effects of feedback on students performance (; Wisniewski et al., 2020; Brooks et al., 2019), with researchers examining variables that may explain the effectiveness of feedback. In addition to characteristics of feedback message (Winstone et al., 2017), mode of delivery (Lyster and Saito, 2010), or context (Gielen et al., 2010), a number of student variables explain variation in differential effects of feedback on achievement. Self-efficacy, prior achievement, emotions, motivation, among others, have been linked to students' processing of feedback and performance improvements (Winstone et al., 2017; Lipnevich & Smith, 2009). To date, there are no studies that have discussed receptivity to feedback as a disposition and that examined its links to grades or other educational attainment indicators. It is safe to presume, however, that students who value feedback, enjoy receiving it, understand it, and know what to do with it, would be more inclined to invest effort into feedback implementation thus contributing to their improved outcomes. We intend to examine this contingency in the current study.

Further, studies have consistently revealed gender effects in various characteristics predicting educational outcomes (e.g., anxiety, Goetz et al., 2013; self-efficacy, Huang, 2016), and to date, no study has examined gender differences in feedback receptivity, primarily because of the recency in the contruct development. In order to make accurate and meaningful comparisons of mean receptivity scores across gender groups, measurement equivalence of the scales must be established (Drasgow & Guertler, 1987; Drasgow & Kang, 1984). Hence, we aimed at examining the equivalence of a measure of receptivity between gender groups to assure that the same construct is being assessed in each group.

The Current Study

The purpose of the current study is threefold. First, we report on the validation of the instrument to measure receptivity to instructional feedback and provide validity evidence for its use, downward extending its original sample from college students to secondary school students. Specifically, we examine generalizability of the construct to a sample of secondary school students from Singapore (AERA, APA, NCME, 2014; ITC, 2016; Wu et al., 2016). Second, we tested a nested hierarchy of hyphotesis for addressing the cross-group (i.e., gender) invariance of the instrument's psychometric propierties and compared means on RIF sub-scales between gender groups. Third, we examined whether sub-scales of RIF predicted student grades.

To this end, we formulated the following research questions:

1. To what extent is there evidence to support structural validity of the Receptivity to Instructional Feedback (RIF) scale?
2. Does receptivity to instructional feedback instrument exhibit adequate cross-gender equivalence? What are the mean diffences in RIF scale scores between gender groups?
3. Do the components of receptivity to feedback at baseline predict grades at subsequent points of data collection?

Method

2.1 Participants and procedure

Participants in this study were *N* = 314 secondary students from Singapore enrolled in five schools. These schools were drawn from a representative range of schools: one was an autonomous school (similar to U.S. private schools), three government and one government-aided schools (similar to U.S. public schools). In each school, three classes of secondary schools were involved. The classes were three Normal Technical classes, three Normal (Academic) classes and nine Express classes. Among the participants, 54.1% (n = 170) self-identified as girls and 45.9% (n = 144) as boys.

The data were collected online, with the exception of one school, where the teacher distributed surveys in the paper-and-pencil format. In the latter case the data were entered by the research assistant. The data were collected at baseline and then in three waves, with the average of three weeks elapsing between baseline and wave 1 as well as wave 1 and wave 2. The distance depended upon each school's writing schedule. Between waves 2 and 3 all schools switched into home-based learning, so the writing task was deffered. The schools resumed one month later, when wave 3 data were collected.

Application for the ethics clearance was submitted to Nanyang Technological University Institutional Review Board (NTU-IRB) prior to the start of data collection. Consent forms, information sheets, study procedures, as well as instruments received approval from the NTU-IRB. Students and their parents gave consent to participate in the study by signing the same form, and the form was accompanied by an information sheet which was also approved by the institutional review board.

2.2 Instrumentation

2.2.1 Receptivity to Instructional Feedback (RIF). The Receptivity to Instructional Feedback (RIF) scale is a self-report instrument designed to measure students' acceptance of instructional feedback. The scale has been explored in a sample of the US and New Zealand university students (Lipnevich et al., 2021). The items were slightly modified to reflect secondary school context (e.g., “professor” was replaced with “teacher”). A total of 36 Likert-type items measured on a 5-point scale (1=strongly disagree and 5=strongly agree) was generated under four receptivity components: (1) experiential attitudes towards feedback, or affective engagement with feedback (e.g., I look forward to receiving the instructor's comments on my work); (2) instrumental attitudes towards feedback (i.e., value for feedback; e.g., I find the comments I get on my assignment to be very helpful); (3) cognitive engagement with feedback (e.g., I know how to use feedback comments to improve my work); and (4) behavioural engagement (e.g., When I receive feedback, I carefully read every comment). Additionally, behavioural engagement with the feedback scale was measured in three subsequent points of data collection.

2.2.2 Grades

Student baseline marks were provided by the teachers using their year-end results from the previous year. The subsequent (wave 1-wave 3) grades were also provided by the teachers after each assignment. Grades were expressed as percentages and descriptive statistics (see Table 1) indicate that the lowest mean on grades was on baseline (M=61.895, sd=8.854, n=166) and the highest on wave 1 (M=65.023, sd = 8.454, n=183), followed by wave 3 (M=64.317, sd=9.042, n=249).

2.3 Analytic Plan

To answer the first research question, a Confirmatory Factor Analysis was conducted to verify the 4-latent factor structure (e.g number of factors, pattern of loading, and correlations among factors) of the RIF scale proposed by Lipnevich et al. (2021). Full-information data was analysed throught Structural Equation Modelling (SEM) using *lavaan package in R* (Rosseel, 2012) with weighted least squares mean and the variance adjusted (WLSMV) estimator, which is a robust estimation method designed explicitly for ordinal data (Sass et al., 2014). Items that had a factor loading of $\lambda \leq 0.5$ were excluded from the corresponding scale or factor, and model alterations at the indicator level were conducted to improve model fit using several model iteractions. Furtheremore, the overall model fit for measurement analyses was evaluated using different indices (Cheung & Rensvold, 2002; Fan & Sivo, 2005, 2007). We used the following indices and their cut-offs for ‘acceptable’ or ‘good’ fit (Brown et al., 2014; Browne & Cudeck, 1992; Hu & Bentler, 1998, 1999; MacCallum et al., 1996; Yu, 2002; Hair, et al., 2010): (1) the Root Mean

Table 1

Descriptive Statistics of Study Variables

	N	Mean	SD	Median	Min	Max	Range	Skewness	Kurtosis
Receptivity of Instructional Feedback									
Behavioral Engagement	307	-0.001	0.945	-0.030	-2.219	2.362	4.581	0.358	-0.169
Cognitive Engagement	307	0.000	0.889	0.091	-2.413	1.936	4.348	0.036	-0.172
Experiential Attitudes	307	0.000	0.911	-0.016	-2.216	1.841	4.057	0.199	-0.525
Instrumental Attitudes	304	-0.001	0.944	0.068	-2.416	1.788	4.204	0.152	-0.494
Behavioral Engagement									
Behavioral Engagement - Wave 1	294	-0.001	0.956	0.002	-3.338	2.091	5.429	0.008	0.002
Behavioral Engagement - Wave 2	299	-0.001	0.945	-0.046	-2.853	2.361	5.214	0.150	-0.016
Behavioral Engagement - Wave 3	291	-0.001	0.948	-0.056	-2.518	2.108	4.626	0.128	-0.264
Grades									
Grades - Baseline	166	61.895	8.854	62.857	34.286	81.000	46.714	-0.293	-0.358
Grades - Wave 1	183	65.023	8.454	65.000	43.330	83.330	40.000	-0.158	-0.514
Grades - Wave 2	152	62.478	12.447	63.333	0.000	96.667	96.667	-1.382	4.593
Grades - Wave 3	249	64.317	9.042	63.333	26.667	100.000	73.333	-0.385	1.793

Square Error of Approximation (RMSEA) with values < .08 being indicative of reasonable fit and values < .05 indicating a good fit; (2) the Comparative Fit Index (CFI) and Tucker-Lewis index (TLI) with values > .90 indicating an acceptable fit and values > .95 indicating a good fit; and (3) the standardized root mean square residual (SRMR) with values < .05 being indicative of good fit.

Scores for each latent variable were estimated using an Item Response Theory (IRT) model. IRT allows for the examination of the scale quality by checking the extent to which items within a scale reflect a single underlining unidimensional latent construct (Kolen & Brennan, 2014; Shevlin et al., 1997, Lu et al., 2005). We used a Graded Response Model (GRM), which is appropriate for the polytomous and ordinal nature of the items with k categories of responses[1]. This model assumes that each factor can be expressed as a latent score (θ). The relationship between each item and θ is described by a slope parameter (discrimination or a-parameter) and one or more location parameters (difficulty or b-parameter). The slope parameter is interpreted as an item's ability to discriminate among different levels on the θ, whereas the location parameters indicates the trait level necessary (θ) to have a 50% probability of choosing the category k or higher as a response (Toland, 2014, Jessen et al., 2018, Bean & Bowen, 2021).

We examined various indices to assess model adequacy. For evaluating the absolute fit of the model to each item we used the generalized S-χ^2, recommended for polytomous data (Orlando & Thissen, 2000, 2003). The generalized S-χ^2 tests the difference between empirical (observed) and model-predicted responses by item response category (Bean & Bowen, 2021; Toland, 2014;

---

[1] Technical manual containing R code, data, codebook, and additional information for scoring Receptivity to Instructional Feedback scales is available in https://osf.io/5xnz7/ (Lipnevich & Lopera-Oquendo, 2022).

Xu et al., 2017). For overall model fit, we used the M2 limited information goodness-of-fit statistics and the associated RSMEA index (Cai & Hansen, 2013; Maydeu-Olivares, 2005; Maydeu-Olivares & Joe, 2014; Toland, 2014). In the context of IRT Models, well-fitting statatistics will have a nonsignificant p-value with a small RSMEA value (RMSEA ≤ 0.089 for adequate fit and RMSEA ≤ 0.05 for close fit (Maydeu-Olivares & Joe, 2014)[2].

For testing the factor invariance hypotheses, multisample and 4-factor CFA models for RIF scale by gender were fit to data using weighted least squares mean estimation (WLSMV). Multisample CFA models fit to the data included the configural, metric, strong, partial strong, and partial strict factorial invariance models as well as equal variance and covariance factor models (Gregorich, 2006; Lugtig et al., 2012; Milfont & Fischer, 2010; Millsap & Yun-Tein, 2004; Schoot et al., 2012; Vandenberg & Lance, 2000)[3]. We evaluated the overall nested model fit using different Comparative Fit Index (CFI) and Tucker-Lewis index (TLI) and absolute indices (RMSEA and SRMR). Fit is considered adequate if the CFI and TLI values are > .90, and better if they are > .95. TLI can become > 1.0, which can be interpreted as an indication of over fitting, making the model more complex than needed. The cut-off value is RMSEA is .08, but RMSEA and SRMR below .05 are commounly considered rough indicadores of good model fit. Aditionally Satorra -Bentler (SB) scaled χ^2 statistics ($\Delta\chi^2$) and differences in model degrees of freedom (Δdf) were computed for testing wheter the more constrained model result in significant worsening of fit (Chen, 2007; French & Finch, 2006). Finally, considering that the outcome

[2] For scales with a small number of items it is suggested to evaluate the item fit statistics at the 1% of confidence level (Stone & Zhang, 2003)

[3] An invariance analysis of the instrument across the two types of instrument administration (online vs. paper-and-pencil) was also conducted (see Tables S9 and S10, Suplemmentary Material). Findings support measurement (i.e., configural, metric, strict, variance and covariance invariance) equivalence of the RIF by the type of administration. So, both types of data sources were included in the analysis.

variable (grades) is given in what is called events/trials form that record the number of successes y_i that occurs in n_i trials (total of possible mark), we used the y_i/n_i (proportion of correct marks) to estimate the probability of success π_i associate with a set of exploratory variables (e.g. gender and receptivity to instructional feedback). Finally, for addressing the third research question, we conducted a logistic regression model with a logit link function, dealing with the restrictions of nonlinearity in our dependent variable (Friendly et al., 2015; H.-G. Müller & Stadtmüller, 2005; M. Müller, 2004).

Results

3.1 Measurement of Receptivity to Feedback

In order to examine research question 1, we evaluated goodness-of-fit of alternative models to understand and provide validity evidence for the factor structure of the RIF (Noar, 2003; Strauss & Smith, 2009), scale based on Lipnevich et al. (2021). Lipnevich et al. (2021) showed that RIF scale had a 4-latent factor structure, with factors representing the initially hypothesized theory-based structure of the measure. Therefore, we compared three CFA models. Table 2 provides details of the model fit indices of the measurement model iterations. Model 1 included all 36 items distributed in the original 4-factors (e.g. Behavioral Engagement, Experiential Attitudes, Instrumental Attitudes, and Cognitive Engagement). Based on absolute indices (RMSEA = 0.103 and SRMR =0.073), the model showed an inadequate fit (see, Table 2). Then, items that insufficiently represented the designated construct (with lower loading factors ($\lambda \leq$ 0.5), were excluded from the subsequent model iteration to improve overall model fit (Frohlich, 2002; Voss et al., 2003) [4]. Model 2, run with 31 items, showed a poor fit according to absolute

[4] Three items from Behavioral Engagement were excluded "I only look at feedback quickly" (λ = 0.267), "I ask my teacher to explain comments I do not understand" (λ = 0.485), "I spend a lot of time studying teacher's comments"

indices (RMSEA = 0.103, and SRMR = 0.073). Finally, we examined Model 3, which contained a reduced set of 24 items proposed by Lipnevich et al. (2020). CFA analyses revealed good fit for Model 3: RMSEA = 0.061 (90% CI: 0.053, 0.068), CFI = 0.992, TLI = 0.991, and SRMR = 0.059. Previous results also showed that scales could be improved by eliminating some of the items without significantly altering the content of the scale, which has important implications for administration of the instrument in classroom settings, especially in middle and high school contexts. Table S1 (Supplementary Material) shows all original items of the scale and their final factor loading to their respective RIF indicators. The correlation between scales in Model 3 ranged from $.688 < r < .791$ (see, Table S2 Supplementary Materials), whereas the internal consistency reliability statistics across the 4 scales ranged from $.74 < \alpha < .86$ (see, Table S3 Supplementary Materials).

($\lambda = 0.476$), while two items were deleted from Experiential attitude factor “I hate it when the teacher hands back the work I have done” ($\lambda = 0.485$), and “I hate it when the teacher hands back the work I have done” ($\lambda = 0.304$)

Table 2

Model Fit Statistics of Factor Analyses Measurement Model Iterations of the Receptivity to Instructional Feedback (RIF) Scale

Models	DVs	Model specification	N	*k*	χ^2	*df*	RMSEA	(90 % CI)	CFI	TLI	SRMR
CFA Model 1	36	4-factor	307	186	2476.02	588	0.104	(0.093, 0.107)	0.972	0.970	0.081
CFA Model 2	31	4-factor	307	161	1784.74	428	0.103	(0.098, 0.108)	0.979	0.977	0.073
CFA Model 3	24	4-factor	307	126	516.10	246	0.061	(0.053, 0.068)	0.992	0.991	0.059

Note. All of the CFA models contained 4 latent factors, each representing the theory-based 4-factor structure of the measure. The first CFA model, Model 1, included all original 36 items of the measure estimated onto their respective hypothesized constructs.

Furthermore, for evaluating the factor structure of the Behavioural Engagement scale in subsequent point data collection (wave 1 thru 3), two CFA models were estimated. The first CFA Model included all original 12 items by scale-wave, whereas Model 2 (i.e., the final measurement model) contained a reduced set of 9 items corresponding with the structure suggested for this factor according to the results from the complete RIF instrument. Model 2 had good fit: RMSEA = 0.0785 (90% CI: 0.072, 0.849), CFI = 0.986, TLI = 0.984, and SRMR = 0.067. The internal consistency reliability statistics (α) for the scale were 0.887, 0.844 and 0.860 for waves 1 through 3 (See, Table S3 Supplementary Material). Tables S4 and S5 (Supplementary Material) show the factor loading and details of the model fit indices of the measurement model iterations, respectively.

Concerning the scoring procedure, the results of the item parameters obtained with the GRM model are shown in Tables S6 and S7 (Supplementary Materials). Overall, the items for RIF scales covered a wide range of latent traits[5]. Regarding a-parameter (discrimination), items had values that suggested a high capacity of response category to distinguish among different latent traits. The individual assessment of each item to the GRM indicated that four items did not fit the **S**-χ^2 p-value $\geq$ 0.01[6]; however RMSEA-χ^2 is a lower cut-off criterion in all cases. Additionally, the global assessment of RIF scales was adequate (See Table S8, Supplementary Material). Cognitive Engagament showed a higher RMSEA, which could be a consequence of this test can be overly sensitive to small model-data misfit (Toland, 2014).

---

[5] Range of b-parameter covers the following ranges by scales: Experiential Attitudes from -6.113 (b_1 item 4) to 1.505 (b_4 item 6), Instrumental Attitudes from -3.545 (item 5) to 1.014 (b_4 item 5), Cognitive Engagement from -4.205 (b_1 item 3) to 1.421 (b_4 item 4) and Behavioral Engagement from -4.120 (b_1 item 8) to 1.679 (b_4 item 7)

[6] "*I look forward to receiving the teacher's comments on my work*" (Experiential Attitudes), "*Feedback on tests and assignments doesn't help me very much*" (Instrumental Attitudes), "*I work through the feedback I receive*" (Behavioral Engagement)

3.2 Invariance analysis and mean differences

The first step in establishing measurement invariance for the four-latent factor structure suggested by RIF instrument was configural invariance. We tested whether the same factorial structure held in both group (boys and girls) separately, without any equality constraints (Vandenberg et. al, 2000). Table 3 shows the results of models fit for measurement invariance. Model 1 had a good fit: RMSE = 0.039, (09% CI=0.027 - 0.049), CFI =0.947, TLI = 0.940, and SRMR= 0.067. Then, we proceeded to testing metric invariance. Metric invariance is a constrained version of configural model where the factor loading are assumed to be equal across groups but the intercepts are allowed to vary between groups. This implies that respondents across gender attribute the same meaning to the latent constructs of RIF. To test metric invariance, we needed to compare the configural model against the metric model using a chi-square difference ($\Delta \chi^2$) test. Table 4 shows that the chi-square difference test was not statistically significant ($\Delta\chi^2 = 25.837$, $df = 20$, $p = 0.1713$). This finding suggests that after constraining the factor loadings to be equal across groups, the model fit did not change substantially. In other words, the constrained model (i.e., Model 2 full metric invariance) fit the data as well as the free model (Model 1). The model fit indices also indicate a good fit for the metric model (RMSE = 0.036, (09% CI=0.023 - 0.046), CFI =0.953, TLI = 0.949, and SRMR= 0.078).

Table 3

Comparison of Fit Indices in Invariance Models by Gender

Model	χ^2	df	χ^2/df	RMSEA	(90 % CI)	CFI	TLI	SRMR
Model 1. Full Configural Invariance	603.96	492	1.228	0.039	(0.027 - 0.049)	0.947	0.940	0.067
Model 2. Full Metric Invariance	611.32	512	1.194	0.036	(0.023 - 0.046)	0.953	0.949	0.078
Model 3. Full Scalar Invariance	641.92	532	1.207	0.037	(0.025 - 0.047)	0.948	0.946	0.080
Model 4. Partial Scalar Invariance	626.04	524	1.195	0.036	(0.024 -0.046)	0.952	0.949	0.078
Model 5. Strict Invariance	654.60	548	1.195	0.036	(0.024 - 0.046)	0.950	0.949	0.082
Model 6. Variance and Covariance Invariance	614.92	558	1.102	0.026	(0.003 - 0.038)	0.973	0.973	0.084

Table 4

Nested Invariance Models Comparison by Gender

Model Comparison	$\Delta\chi^2$	Δdf	ΔCFI	ΔTLI	ΔRMSEA	ΔSRMR	Decision
Configural vs Metric Invariance	25.837	20	0.006	0.009	-0.003	0.011	Accept
Metric vs Scalar Invariance	45.36	20***	-0.005	-0.003	0.001	0.002	Reject
Metric vs Partial Scalar Invariance	19.501	12	-0.001	0.000	0.000	0.001	Accept
Partial Scalar vs. Strict Invariance	36.592	24	-0.002	0.000	0.000	0.004	Accept
Strict vs. Variance and Covariance Invariance	7.256	10	0.024	0.024	-0.01	0.002	Accept

Note *** p.value < 0.001, ** p.value < 0.01, * p.value < 0.05, . p.value < 0.1

After metric invariance was established, the next step was to impose scalar invariance, a constrained version of the metric model where both the factor loadings and intercepts are assumed to be equal between girls and boys. Scalar invariance implies that the meaning of the construct (the factor loadings), and the levels of the underlying items (intercepts) are equal in both groups. Therefore, observed scores are related to the latent score; that is, individuals who have the same scores on the latent construct would obtain the same score on the observed variable regardless of their group membership (Schoot et al., 2012; Vandenberg & Lance, 2000).

Although model fit indices for full scalar invariance (Model 3) indicated an acceptable fit (RMSE = 0.037, (09% CI=0.025 - 0.047), CFI =0.948, TLI = 0.946, and SRMR= 0.080), the statistically significant results suggested that there was a lack of scalar invariance by gender for the RIF instrument ($\Delta\chi^2 = 45.36$, $df = 20$, $p < 0.0001$). Consequently, we tried to established partial measurement invariance, which may allow appropriate cross-group comparison even if full measurement invariance was not obtained (Byrne et al., 1989; Steenkamp & Baumgartner, 1998).

The goal of testing for partial measurement invariance is to find out which of the loadings or intercepts differ across groups. If only one of these is different across groups, we know that any differences on the latent variable can either be caused by a difference in this loading/intercept, or by the true latent variable group difference. Therefore, we identified eight items[7] that had a significant impact on model fit, due to parameters not being invariant across group, and instead of fixing these intercept parameters, we estimated them freely for girls and

[7] Items identified with influential parameters by scale were: i) Experiential Attitutes ("*I enjoy reading teacher's comments on my tests/assignments*", and "*I do not like it when my work is given a mark*"), ii) Cognitive Engagement ("*The comments the teacher makes on my work are easy to understand*"), iii) Behavioral Engagement ("*I go over teacher's comments several times*", "*When I receive feedback, I think about how I would do things differently next time*", "*I don't really think about the feedback I receive*", "*I don't really think about the feedback I receive*"), and iv) Instrumental Attitudes ("*Teacher's feedback is very effective in helping me enhance my performance*").

boys. That is, we released the constraints for these parameters to establish scalar partial measurement invariance (Model 4). The comparison of the adjusted scalar model and metric model indicated that the chi-square difference test was not significant ($\Delta\chi^2 = 19.501$, *df* = 12, *p* = 0.077). Also, the model fit showed that the adjusted partial scalar model had a good fit (RMSE = 0.036, (90% CI=0.024 -0.046), CFI =0.952, TLI = 0.949, and SRMR= 0.078).

Next, we tested strict factorial invariance, a constrained version of the scalar model where the factor loadings, intercepts, and residual variances are fixed across groups. The model of partial strict invariance (Model 5) was compared to the partial scalar invariance model (Model 4). The difference in terms of chi-squeared was significant ($\Delta\chi^2 = 36.59$, df = 24, p = 0.048), but the CFI decrease was trivial -0.002, whereas SMRS improved (0.004). Hence, it can be concluded that partial strict invariance was supported (Steenkamp & Baumgartner, 1998). Finally, we tested the invariance of factor variance and covariance (Model 6). Invariance of factor variance indicated that the range of scores on a latent factor did not vary across groups, whereas the factor covariance invariance tested if all latent variables had the same relationship in all groups. This model was compared with partial strict invariance (Model 5). The difference in terms of chi-squared was not significant ($\Delta\chi^2 = 7.256$, df = 10, p = 0.7011). Furthermore, the model fit indices for the variance and covariance invariant model showed a good fit (RMSE = 0.026, (90% CI=0.003 - 0.038), CFI =0.973, TLI = 0.973, and SRMR= 0.084). Thus, the invariance variance and covariance was supported.

After we established invariance, we conducted independent samples t-tests to determine whether there were significant differences in means of RIF scales across gender groups. Overall, girls had had higher scores on RIF scales than boys (see Table S11 Supplementary Material), with behavioral engagement (.295, p-value <0.01) and experiential attitudes (.292, p-value <0.01) reaching statistical significance.

3.3 Instructional feedback and grades

Logistic regression models (see Table 5), controlling for gender, examined the degree to which each receptivity of feedback scale predicted the probability to obtain higher marks at baseline and subsequent data points. Results across the regression analyses in the four receptivity measures showed that the probability of getting higher marks is lower for males than females. Increases in experiential attitudes and cognitive engagement were significantly and positively associated with an increase in the likelihood of higher marks in the baseline and wave 1. In contrast, instrumental attitudes predicted increments in grades in baseline and wave 3. Furthermore, behavioral engagement in wave 1 predicted the probability of obtaining higher marks (grades) in wave 2, whereas behavioral engagement in waves 2 and 3 was associated within marks in corresponding waves (see Table 6). Therefore, the results suggested evidence for discriminant validity among the receptivity factors and their relevance for predicting meaningful educational outcomes. However, the predictive power of cognitive engagement on grades[8] must be carefully analyzed because even though individual items showed a good fit to the scoring model, the overall fit of the GRM model was not adequate according to the traditional cut-off criterion, suggesting that effects could be biased.

[8] Models using cognitive engagement scoring with CFA were also estimated and results were similar to those provided with IRT scores.

Table 5

Logistic Regression. Receptivity to Instructional scales in Baseline as a predictor of Grades

Variable	BL			Wave 1			Wave 2			Wave 3		
	β	OR	SE	β	OR	SE	β	OR	SE	β	OR	SE
Model 1												
Intercept	0.637***		0.02	0.732***		0.04	0.708***		0.05	0.676***		0.03
Gender (Male)	-0.346***	0.708	0.04	-0.265***	0.767	0.06	-0.315***	0.729	0.07	-0.245***	0.783	0.05
Behavioral Engagement	0.018	1.018	0.02	0.040	1.041	0.03	0.011	1.011	0.03	0.035	1.036	0.03
Model 2												
Intercept	0.638***		0.02	0.726***		0.04	0.707***		0.05	0.678***		0.03
Gender (Male)	-0.349***	0.705	0.04	-0.265***	0.767	0.06	-0.315***	0.730	0.07	-0.250***	0.779	0.05
Cognitive Engagement	0.041*	1.042	0.02	0.073*	1.076	0.03	0.023	1.023	0.04	0.035	1.036	0.03
Model 3												
Intercept	0.637***		0.02	0.721***		0.04	0.704***		0.05	0.679***		0.03
Gender (Male)	-0.339***	0.712	0.04	-0.258***	0.773	0.06	-0.309***	0.734	0.07	-0.252***	0.777	0.05
Experiential Attitude	0.038*	1.038	0.02	0.070*	1.072	0.03	0.026	1.027	0.04	0.019	1.019	0.03
Model 4												
Intercept	0.642***		0.02	0.731***		0.04	0.707***		0.05	0.675***		0.03
Gender (Male)	-0.349***	0.705	0.04	-0.271***	0.763	0.06	-0.309***	0.734	0.07	-0.248***	0.780	0.05
Instrumental Attitudes	0.036*	1.037	0.02	0.040	1.040	0.03	0.028	1.028	0.04	0.055*	1.057	0.03
N	162			180			146			243		

Note . p <0.1 =, * p < 0.05; ** p < 0.01, *** p < 0.001.

Table 6

Logistic Regression. Behavioral Engagement in Wave1 to 3 as a predictor of Grades

Variable	Grades								
	Wave 1			Wave 2			Wave 3		
	β	OR	SE	β	OR	SE	β	OR	SE
Model 5									
Intercept	0.726***		0.04	0.682***		0.05	0.680***		0.03
Gender (Male)	-0.249***	0.779	0.06	-0.260***	0.771	0.07	-0.219***	0.803	0.05
Behavioral Engagement - Wave 1	0.041	1.042	0.03	0.075*	1.078	0.04	0.010	1.010	0.03
N	175			140			234		
Model 6									
Intercept				0.712***		0.05	0.676***		0.03
Gender (Male)				-0.266***	0.767	0.07	-0.203***	0.816	0.05
Behavioral Engagement- Wave 2				0.066 .	1.069	0.04	0.029	1.029	0.02
N				143			238		
Model 7									
Intercept							0.673***		0.03
Gender (Male)							-0.255***	0.775	0.05
Behavioral Engagement - Wave 3							0.045.	1.046	0.03
N							240		

Note . p <0.1 =, * p < 0.05; ** p < 0.01, *** p < 0.001.

Discussion

In the current study we collected validity evidence for the instrument intended to measure receptivity to instructional feedback. Specifically, we downward extended the sample from the university (Lipnevich et al., 2021) to the secondary school students. We explored generalizability of the instrument to a different culture (i.e., Singapore) and examined cross-gender stability of the construct (AERA, APA, NCME, 2014; ITC, 2016; Wu et al., 2016). To achieve the latter goal, we tested a nested hierarchy of hyphotesis for addressing the cross-group (i.e., gender) invariance of the instrument's psychometric propierties. This allowed us to examine differences in mean receptivity scores between boys and girls. Finally, we examined whether sub-scales of RIF predicted student grades.

The CFA results confirmed the existence of four separate factors of receptivity to feedback: experiential attitudes toward feedback, instrumental attitudes toward feedback, cognitive engagement with feedback, and behavioural engagement with feedback. Hence, the structure established in a sample of the US and New Zealand university students held for secondary students in Singapore.

Further, no previous study has examined the measurement equivalence of receptivity across any demographic groups and failure to establish measurement equivalence between gender groups would have had practical implications for the interpretations of the receptivity score. We found support for measurement (i.e., configural, metric, partial scalar, strict, and variance and covariance invariance) equivalence of the RIF. Because of the equivalence of the factor coefficients and intercepts, it was possible to make meaningful comparisons between boys and girls. We found that girls' scores on experiential attitudes and behavioral engagement scales were significantly higher than boys'. When considering the Cohen's d-values, which is an expression of the difference in standard deviation units, the difference was .315 and .324 of one standard deviation for behavioral engagement and experiential attitude, respectively. In other words, girls view feedback more favorably than boys and like receiving feedback on their assignments, and they also report to employ a larger range of behaviors upon receiving feedback.

Our study also demonstrated that receptivity to feedback made substantial contributions to predicting student grades. So, after controlling for gender, behavioral engagement and experiential attitudes explained 15% of variance in student grades. Conscientiousness and self-efficacy have been consistently found to be the strongest personality predictors of student academic attainment. So, Poropat, (2009) conducted a meta-analysis and revealed an average correlation of $r = 0.22$ between conscientiousness and grades. Other studies have shown that conscientiousness accounted for an additional 10% of the variance in GPA even after controlling for intelligence (Di Fabio & Busoni, 2007) and conscientiousness, extraversion, and openness to experience explained additional 12% of the variance in grades of college students (Furnham & Chamorro-Premuzic, 2004). Furthermore, Robbins et al. (2003) showed that the best predictors for GPA were students' academic self-efficacy and achievement motivation with average correlations of r =.496 and .303. Hence, the contribution of receptivity to explaining student grades is noteworthy. In terms of its practical value, one of the most important implications of this study is that receptivity may represent a relatively malleable characteristic, and hence, may be influenced through instructional interventions (Winstone et al., 2017). For example, enhancing the perceived value of feedback (instrumental attitudes) and teaching student specific strategies for feedback uptake (behavioral engagement) could be relatively accessible and highly instrumental in their use of feedback. Future research should examine RIF's incremental prediction of grades over and above personality, self-efficacy, and motivation.

This study is not without limitations. Due to the specifics of study design, the reported study could not be used to establish causal relationship between receptivity and performance. Future studies could employ a longitudinal designs to examine links between receptivity and achievement. Further, other factors may have influenced the magnitude of the relationship between receptivity and grades. Hence, exploring receptivity and its links to indicators of academic performance could be a fruitful area for future research. Although the purpose of the study was not to examine, in detail, the IRT model, we offered a collection of statistical models to define the relationship between individuals' unobserved latent scores and item characteristics of the RIF scale. Future studies may further examine IRT assumptions and hypotheses through systematic variations in the data structure, such as modeling scores using multidimensional IRT. Further work on exploring generalizability of the RIF to other contexts, educational levels, and countries is also in order.

Conclusion

Our findings revealed that the factor structure of the receptivity measure was maintained in a sample of secondary school students from Singapore. Furthermore, measurement equivalence was established for the scores of the four factors of the RIF scale between gender groups. That is, RIF items measuring the latent receptivity to feedback facets are interpreted the same way across male and female secondary school students. Mean comparison revealed higher scores on experiential attitudes and behavioral engagement for girls. Furthermore, controlling for gender, experiential attitudes (or affective engagement), cognitive, and behavioral engagement predicted student grades. Hence, the measure is appropriate for usage in a secondary school sample in Singapore and is useful for predicting meaningful educational outcomes. In sum, receptivity to feedback may be highly useful in predicting grades, and its malleability may be explored through class-wide interventions and instructional activities. Effective and psychometrically sound tool will be helpful to achieve this goal.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. American Educational Research Assn.

Bean, G. J., & Bowen, N. K. (2021). Item Response Theory and Confirmatory Factor Analysis: Complementary Approaches for Scale Development. *Journal of Evidence-Based Social Work*, *18*(6), 597–618. https://doi.org/10.1080/26408066.2021.1906813

Brown, G., Harris, L., & Harnett, J. (2014). Understanding classroom feedback practices: A study of New Zealand student experiences, perceptions, and emotional responses. *Educational Assessment, Evaluation and Accountability*. https://doi.org/10.1007/s11092-013-9187-5

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *The British Journal of Mathematical and Statistical Psychology*, *66*(2), 245–276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Di Fabio, A., & Busoni, L. (2007). Fluid intelligence, personality traits and scholastic success: Empirical evidence in a sample of Italian high school students. *Personality and Individual Differences*, *43*(8), 2095–2104. https://doi.org/10.1016/j.paid.2007.06.025

Drasgow, F., & Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology*, *72*(1), 10–18. https://doi.org/10.1037/0021-9010.72.1.10

Drasgow, F., & Kang, T. (1984). Statistical power of differential validity and differential prediction analyses for detecting measurement nonequivalence. *Journal of Applied Psychology*, *69*(3), 498–508. https://doi.org/10.1037/0021-9010.69.3.498

Fan, X., & Sivo, S. A. (2005). Sensitivity of Fit Indexes to Misspecified Structural or Measurement Model Components: Rationale of Two-Index Strategy Revisited. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 343–367. https://doi.org/10.1207/s15328007sem1203_1

Fan, X., & Sivo, S. A. (2007). Sensitivity of Fit Indices to Model Misspecification and Model Types. *Multivariate Behavioral Research*, *42*(3), 509–529. https://doi.org/10.1080/00273170701382864

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*(3), 378–402. https://doi.org/10.1207/s15328007sem1303_3

Friendly, M., Meyer, D., & Zeileis, A. (2015). Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data. In *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data* (p. 525). https://doi.org/10.1201/b19022

Frohlich, M. T. (2002). Techniques for improving response rates in OM survey research. *Journal of Operations Management*, *20*(1), 53–62. https://doi.org/10.1016/S0272-6963(02)00003-7

Furnham, A., & Chamorro-Premuzic, T. (2004). Personality and intelligence as predictors of statistics examination grades. *Personality and Individual Differences*, *37*(5), 943–955. https://doi.org/10.1016/j.paid.2003.10.016

Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2013). Do Girls Really Experience More Anxiety in Mathematics? *Psychological Science*, *24*(10), 2079–2087.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(11 Suppl 3), S78-94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, C. (2016). Achievement goals and self-efficacy: A meta-analysis. *Educational Research Review*, *19*, 119–137. https://doi.org/10.1016/j.edurev.2016.07.002

Jessen, A., Ho, A. D., Corrales, C. E., Yueh, B., & Shin, J. J. (2018). Improving Measurement Efficiency of the Inner EAR Scale with Item Response Theory. *Otolaryngology--Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, *158*(6), 1093–1100. https://doi.org/10.1177/0194599818760528

Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education*, *14*(1), 63–76. https://doi.org/10.1177/1469787412467125

Kang, T., & Chen, T. T. (n.d.). *An Investigation of the Performance of the Generalized S-X2 Item-Fit Index for Polytomous IRT Models*. 30.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices, 3rd ed* (pp. xxvi, 566). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-0317-7

Lipnevich, A. A., Lopera-Oquendo, C. (2022, March 28). Receptivity to Instructional Feedback. Technical Manual. https://osf.io/5xnz7/

Lipnevich, A. A., Lopera-Oquendo, C. (2021, October 7). Data, codebooks and supplementary materials for Receptivity to Instructional Feedback: A Validation Study in the Secondary School Context in Singapore. https://osf.io/83xy7/

Lipnevich, A. A., Gjicali, K., Asil, M., & Smith, J. K. (2021). Development of a measure of receptivity to instructional feedback and examination of its links to personality. *Personality and Individual Differences*, *169*, 110086. https://doi.org/10.1016/j.paid.2020.110086

Lipnevich, A., Berg, D., & Smith, J. (2016). The impact of feedback, grades, scores, and comments on students. *Handbook of Human Factors and Social Conditions of Assessment*, 169–185.

Lipnevich, A., & Smith, J. (2009). Effects of Differential Feedback on Students' Examination Performance. *Journal of Experimental Psychology. Applied*, *15*, 319–333. https://doi.org/10.1037/a0017841

Lu, I. R. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in Structural Equation Models: A Comparison With Regression Based on IRT Scores. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(2), 263–277. https://doi.org/10.1207/s15328007sem1202_5

Lugtig, P., Boeije, H. R., & Lensvelt-Mulders, G. J. L. M. (2012). Change? What Change? *Methodology*, *8*(3), 115–123. https://doi.org/10.1027/1614-2241/a000043

MacCallum, R. C., Browne, M. W., Sugawara, H. M., & Wegener, D. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*. https://doi.org/10.1037/1082-989X.1.2.130

Maydeu-Olivares, A. (2005). Further Empirical Results on Parametric Versus Non-Parametric IRT Modeling of Likert-Type Personality Data. *Multivariate Behavioral Research*, *40*(2), 261–279. https://doi.org/10.1207/s15327906mbr4002_5

Maydeu-Olivares, A., & Joe, H. (2014). Assessing Approximate Fit in Categorical Data Analysis. *Multivariate Behavioral Research*, *49*. https://doi.org/10.1080/00273171.2014.911075

Milfont, T., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, ISSN 2011-7922, Vol. 3, Nº. 1, 2010, Pags. 111-130*, *3*. https://doi.org/10.21500/20112084.857

Millsap, R. E., & Yun-Tein, J. (2004). Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research*, *39*(3), 479–515. https://doi.org/10.1207/S15327906MBR3903_4

Müller, H.-G., & Stadtmüller, U. (2005). Generalized Functional Linear Models. *The Annals of Statistics*, *33*(2), 774–805.

Müller, M. (2004). *Generalized Linear Models*. https://doi.org/10.1007/978-3-642-21551-3_24

Murano, D., Lipnevich, A. A., Walton, K. E., Burrus, J., Way, J. D., & Anguiano-Carrasco, C. (2021). Measuring social and emotional skills in elementary students: Development of self-report Likert, situational judgment test, and forced choice items. *Personality and Individual Differences*, *169*, 110012. https://doi.org/10.1016/j.paid.2020.110012

Orlando, M., & Thissen, D. (2000). Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further Investigation of the Performance of S - X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *27*(4), 289–298. https://doi.org/10.1177/0146621603027004004

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322–338. https://doi.org/10.1037/a0014996

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating Model Fit With Ordered Categorical Data Within a Measurement Invariance Framework: A Comparison of Estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167–180. https://doi.org/10.1080/10705511.2014.882658

Schoot, R., Lugtig, P., & Hox, J. (2012). Developmetrics: A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486–492. https://doi.org/10.1080/17405629.2012.686740

Shevlin, M., Miles, J. N. V., & Bunting, B. P. (1997). Summated rating scales. A Monte Carlo investigation of the effects of reliability and collinearity in regression models. *Personality and Individual Differences*, *23*(4), 665–676. https://doi.org/10.1016/S0191-8869(97)00088-3

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research*, *25*(1), 78–107. https://doi.org/10.1086/209528

Strauss, M. E., & Smith, G. T. (2009). Construct Validity: Advances in Theory and Methodology. *Annual Review of Clinical Psychology*, *5*(1), 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Toland, M. D. (2014). Practical Guide to Conducting an Item Response Theory Analysis. *Journal of Early Adolescence*, *34*(1), 120–151. https://doi.org/10.1177/0272431613511332

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Voss, K. E., Spangenberg, E. R., & Grohmann, B. (2003). Measuring the Hedonic and Utilitarian Dimensions of Consumer Attitude. *Journal of Marketing Research*, *40*(3), 310–320. https://doi.org/10.1509/jmkr.40.3.310.19238

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting Learners' Agentic Engagement With Feedback: A Systematic Review and a Taxonomy of Recipience Processes. *Educational Psychologist*, *52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational Measurement for Applied Researchers: Theory into Practice*. Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5

Xu, J., Paek, I., & Xia, Y. (2017). Investigating the Behaviors of M2 and RMSEA2 in Fitting a Unidimensional Model to Multidimensional Data. *Applied Psychological Measurement*, *41*(8), 632–644. https://doi.org/10.1177/0146621617710464

Yu, C. (2002). *Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes*.

Open Science Statement

Conflict of Interest: We have no known conflict of interest to disclose.

Open Science: We report all data exclusions, all data exclusion criteria (if any), whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If I use inferential tests, I report exact p values, effect sizes, and 95% confidence or credible intervals.

Open Data: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including the codebook and supplements, and they can be retrieved from https://osf.io/83xy7/

Open Materials: The information needed to reproduce all of the reported methodology is not openly accessible. The material is available on request from author(s). Technical manual containing all the syntax, data, and additional information for scoring scales used in this study can be retrieved from https://osf.io/5xnz7/ (Lipnevich & Lopera-Oquendo, 2022).

Preregistration of Studies and Analysis Plans: This study was not preregistered.

ORCID
Anastasiya A. Lipnevich https://orcid.org/0000-0003-0190-8689

Anastasiya A. Lipnevich
Department of Educational Psychology
The Graduate Center, The City University of New York
365 5th Avenue, New York, NY, 10016 USA
e-mail: alipnevich@gc.cuny.edu.