LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Measuring Foreign Language Students' Self-Determination: A Rasch Validation Study

Paul Leeming and Justin Harris
Kindai University, Osaka, Japan

Abstract: Self-determination theory has been applied in various educational contexts. Language learning researchers have used factor analysis and structural equation modeling to investigate measures designed for this theory, but Rasch analysis has not been conducted. The Basic Psychological Needs Satisfaction and Frustration Scale (BPNSFS) and the Language Learning Orientation Scale (LLOS) were adapted for a Japanese tertiary educational context and administered to 600 students. Rasch analysis was applied to investigate the dimensionality of the measures and functioning of individual items. Regression analyses were used to examine the relationship between motivation and performance on an English reading and listening test. Results showed that the constructs were largely unidimensional, supporting use of these measures in future research, although problems with individual items suggest that revisions should be made. Regression analyses revealed that external forms of motivation had limited impact on performance, suggesting that teachers should attempt to enhance students' internalized forms of motivation.

Keywords self-determination theory; motivation; Rasch analysis; validity; measurement

We would like to thank the editors and reviewers at *Language Learning* for their feedback and guidance with this paper. We would particularly like to thank Prof. Kormos, the handling editor for this paper, who gave detailed and helpful comments throughout. In addition, we would like to thank Joseph P. Vitta for his help with the statistical analyses.

Correspondence concerning this article should be addressed to Paul Leeming, Kindai University, 3-4-1 Kowakae, Higashiosaka City, Osaka 577–8502, Japan. E-mail: paulleeming@yahoo.com

The handling editor for this article was Judit Kormos.

DOI: 10.1111/lang.12496

Introduction

Although there is no single factor that determines success in second or foreign language learning, many researchers agree that motivation is extremely important (Dörnyei & Ryan, 2015). In English as a foreign language contexts such as Japan, students often have very limited exposure to the second language (L2) and therefore are required to actively seek opportunities to engage with the language, making motivation even more critical. In recent years, L2 acquisition researchers have begun to adopt self-determination theory (SDT; Deci & Ryan, 1985) as a useful framework for understanding students' basic psychological needs and how these needs guide students' own language learning behaviors (Agawa & Takeuchi, 2016; Lou, Chaffee, Vargas Lascano, Dincer, & Noels, 2018; Noels et al., 2019; Noels, Pelletier, Clément, & Vallerand, 2000; Noels, Vargas Lascano, & Saumure, 2019). SDT, which comes from general psychology research on motivation (Ryan & Deci, 2017), posits that humans have basic psychological needs that need to be met in order for them to function optimally.

In order to understand how the theory relates to practice, researchers have developed a number of questionnaires and used them to measure different aspects of SDT in relation to language learning (Noels et al., 2000). Researchers in Japan have developed and validated their own specific measures, investigating how they apply in a Japanese context (Agawa & Takeuchi, 2016; Hiromori, 2006). However, as with other fields, in applied linguistics it is essential that variables be operationalized with attention to reliability and validity through as many of the available methods as possible (Purpura, Brown, & Schoonen, 2015). One such method is Rasch analysis (Bond, Yan, & Heene, 2021), which has been used extensively in the language testing field (Aryadoust, Ng, & Sayama, 2021; Bokander & Bylund, 2020; McNamara & Knoch, 2012) and is increasingly being used in other areas of applied linguistics research, including studies into self-efficacy (Leeming, 2017; Wang, Kim, Bai, & Hu, 2014), anxiety (Apple, 2013), motivation (X. Chen, Lake, & Padilla, 2021), and willingness to communicate (Weaver, 2005). Despite the benefits that Rasch analysis offers compared to more traditional approaches, our search of the literature revealed that it has never been conducted on SDT measures used for language learning research.

Also related to measurement, a shortcoming of many studies that investigate motivation in language learning is the failure to link psychological measures that rely on self-reported data with outcomes. This has been described as the "questionnaire curse" (Al-Hoorie, Hiver, Kim, & De Costa, 2021, p. 6), where self-report questionnaires are used for all of the data gathered in a study,

including the dependent variable, although there are notable exceptions (e.g. Apple, Hill, & Falout, 2020; Hiver & Al-Hoorie, 2020; Yashima, Nishida, & Mizumoto, 2017). Reio (2010) describes this common method variance bias as a serious threat to the validity of quantitative research studies in the social sciences. This paper reports on a study that used Rasch analysis to investigate the dimensionality and item functioning of two measures used within SDT research in language learning, and then examines the relationship between motivation and performance on an English listening and reading test that is used extensively in Japan, for both employment and promotion within companies (Apple et al., 2020; ETS-TOEIC, 2019).

Background Literature

Self-Determination Theory as a Theory of Motivation

SDT is a comprehensive theory of human motivation that has attracted increasing attention since its origins almost 40 years ago (Deci & Ryan, 1985). One of the central theories within SDT is basic psychological needs theory. This posits that humans need to feel a sense of *autonomy*, or control over their own actions, a sense of *relatedness*, or belonging with those around them, and feelings of *competence*, or the belief that they have the ability to successfully complete tasks that they face (Ryan & Deci, 2017); only when all three needs are met are humans able to function effectively. A measure of these needs, the Basic Psychological Needs Satisfaction and Frustration Scale (BPNSFS), has been developed and is translated in multiple languages (B. Chen et al., 2015).

Another theory that is integral to SDT is organismic integration theory (Ryan & Deci, 2017). This posits that there are various kinds of motivation existing on a continuum from *intrinsic motivation*, where people perform a task for the inherent interest and pleasure that it gives them, to various levels of *extrinsic motivation*, in which, in its most controlled form, people complete tasks for some form of external reward, or due to external pressure (see Figure 1). Beyond extrinsic motivation is *amotivation*, where people have no motivation to complete the task. Within this framework, the goal of teachers is to help students to orient toward more internalized forms of motivation, which are considered to be more effective. Organismic integration theory is perhaps the most familiar aspect of SDT for language motivation researchers, and it has been applied in numerous studies, led by the work of Kimberly Noels and her colleagues in North America (Lou et al., 2018; Noels et al., 2019; Noels et al., 2000; Noels et al., 2019). Noels et al. (2000) developed the Language

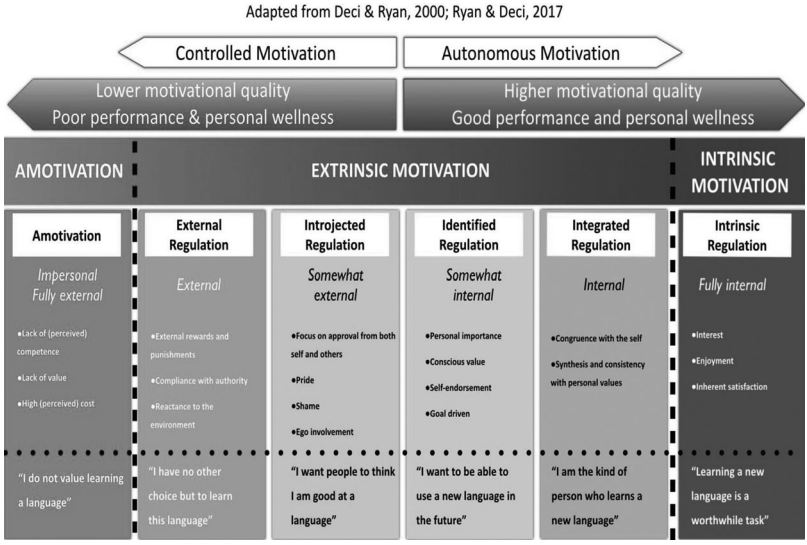

Figure 1 Continuum of motivation in organismic integration theory. From "Self-Determination for All Language Learners: New Applications for Formal Language Education," by M. S. McEown & W. L. Q. Oga-Baldwin, 2019, *System*, *86*, 102124. Copyright McEown & Oga-Baldwin. Reprinted with permission. (Adapted from Deci & Ryan, 2000; Ryan & Deci, 2017)

Learning Orientation Scale (LLOS) to measure the continuum of motivation based on organismic integration theory.

Motivation in the Japanese Context

Motivation is of interest in any educational context, but it has received particular attention in Japan, where universities have been famously described as a "motivational wasteland" (Berwick & Ross, 1989, p. 207). A volume by Apple, Da Silva, and Fellner (2013) was devoted exclusively to motivation research in this context. In an introduction to that volume, Ushioda (2013) argued that motivation is of interest due to the many issues facing English education in Japan and claimed that Japan is a leader in research in this field. The large body of work on "demotivation" in Japan is perhaps also somewhat telling (Arai, 2004; Falout, Elwood, & Hood, 2009; Kikuchi & Sakai, 2009). There are myriad reasons for students becoming demotivated, and Kikuchi (2009) outlined five: the behavior of teachers in the classroom, teaching methodology (grammar–translation), entrance exams, problems related to the rote

memorization of vocabulary, and textbook issues. As well as these, parental pressure may also be a factor. A study by the Benesse Educational Research and Development Institute (2015) of over 5,000 learner–guardian pairs (from the fourth grade of elementary school to the second grade of junior high school) showed that parental expectations of their children's study habits have a noticeable effect on motivation. Using SDT as their motivational framework, the researchers found that children of less "autonomously supportive" guardians (i.e., those who give material rewards for success rather than positive feedback and praise) self-identified as having less intrinsic motivation.

Seemingly aware of motivation issues, the Japanese Ministry of Education has stated it aims to develop students' motivation to learn English (MEXT, 2011), and revisions to the Course of Study (a national outline of teaching curriculum) aim to introduce curriculum changes with that as one of the goals. In an attempt to develop students' interest in the language, since 2020 English is introduced from the third grade of elementary school. This may not adequately address the issue, however, with some research (e.g., Someya, 2018) showing an overall shift from more internalized motivation to external motivation among learners between elementary school and junior high school.

Despite attempts at changing the focus of curricula, during the six years of junior high school and high school, there is a continued focus on studying for English tests in order to be able to enter high school and subsequently university (Apple et al., 2020; Taguchi, Magid, & Papi, 2009; Yashima, Zenuk-Nishide, & Shimizu, 2004). During this time, many students face grueling entrance exams that are based largely on receptive skills. Upon entering university, a vast majority of students are once more faced with a year or more of compulsory English courses. University students, particularly non-English majors with relatively low proficiency, are often motivated only to pass the course and not to achieve higher levels of proficiency in the language (Falout et al., 2009).

Research into SDT and how it can describe Japanese students' motivation to study English has often focused on the development of measures of motivation. In one of the first studies conducted in this context, Yamauchi and Tanaka (1998) used SDT to model the language learning motivation of elementary school students. They concluded that patterns of motivation in Japan mirrored those in other contexts. Later, Hiromori (2006) developed measures of motivation based on SDT specifically for a Japanese context, from which items have subsequently been refined by Agawa and Takeuchi (2016). Using exploratory and confirmatory factor analysis, they found that their model of motivation based on SDT did not fit the data well, with

relatedness and autonomy displaying nonsignificant relationships with other forms of motivation. Researchers in Japan have used SDT to explain motivation in areas such as extensive reading (Tanaka, 2017) and listening (Someya, 2017), and more generally with elementary school students (Oga-Baldwin, Nakata, Parker, & Ryan, 2017; Someya, 2018). Results showed that SDT was successful in modeling the motivation of students. Tanaka (2017) found that autonomy predicted university students' motivation to read, and Someya (2017) showed that intrinsic motivation predicted listening proficiency among junior high school students. Oga-Baldwin et al. (2017) modeled elementary school students' motivation over a year and claimed that SDT was a cohesive and useful framework for understanding motivational change.

A large body of research into motivation has focused on Dörnyei's (2005) L2 motivation self system (L2MSS) model, including many studies in Japan (Apple et al., 2020; Irie & Brewster, 2013; Taguchi et al., 2009; Yashima et al., 2017). The L2MSS model relates to SDT in that the *ought-to self* can be considered to be externalized and less of a positive influence on behavior than the *ideal self*, which represents more internalized forms of motivation. Taguchi et al. (2009) set out to investigate whether studies supporting the model in European contexts (e. g. Dörnyei & Csizér, 2002) could be replicated with students in Japan, China, and Iran. They found that data supported the applicability of the model in those three countries. Yashima et al. (2017) investigated how the L2MSS predicted performance on the Test of English as a Foreign Language (TOEFL) test (https://www.ets.org/toefl/). They found that although ideal selves had a stronger relationship with proficiency, ought-to selves also played a significant role in predicting their outcome variable, counter to studies in other contexts such as Chile and Hungary that found a more limited role for the ought-to self (Kormos & Csizér, 2008; Kormos, Kiddle, & Csizér, 2011). Yashima et al. (2017) argued that the role of the ought-to self is greater in Japan due to a stronger influence on students from other people in their lives, as well as heavy pressure to succeed.

Measurement of Motivation Within Self-Determination Theory

Much research within SDT has focused on the development of measurement instruments, but often with similar methods. Many studies employ Likert scales, assuming that collected data are on an interval scale, and yet Likert category type questionnaires are ordinal (Embertson & Reise, 2000). Researchers commonly add the scores from the Likert category data, assuming that each of the items is contributing equally to the measurement of the construct; for example, it is assumed that a response of "3" on the first item contributes as

much to the measurement of a given construct as a response of "3" on all other items, irrespective of the potential differences in difficulty between items (with Likert-scale category response data, difficulty refers to how easily a statement can be endorsed). Another unfounded assumption is that the difference in Likert categories between a "2" and "3" is the same as that between a "5" and "6" (Wilson, 2005).

Waugh and Chapman (2005) conducted both a Rasch analysis and a factor analysis on the same sets of data (from a social anxiety questionnaire and an attitude to mathematics questionnaire), and whereas factor analysis showed the measures to be reliable, Rasch analysis was able to provide much more detailed information regarding the dimensionality of the measures and also the performance of individual items, revealing serious issues with both. Within the fields of SDT and L2 acquisition, researchers generally develop questionnaires to measure several latent constructs, and interpret the results based on the assumption that each construct is unidimensional and being measured accurately. Without unidimensional measures, the interpretation of results becomes challenging. A limitation of factor analysis is that it depends on correlations from the sample, and there is no measure of fit or standard error for individual items (Wilson, 2005).

The Rasch model (Rasch, 1960) is a one-parameter stochastic model for random probability distribution, showing the relationship between item difficulty and person ability (i.e., for questionnaires, this is an individual's chance of endorsing an item). For questionnaire data, the polytomous Rasch model known as the Rasch rating scale model is used (Andrich, 1978), allowing researchers to measure the thresholds between Likert response categories, such as the difference between *agree* and *strongly agree*, with reference to the level of latent construct that a person has (e.g., level of self-efficacy). This model offers certain benefits for researchers over more traditional forms of validation (Apple & Neff, 2012; Knoch & McNamara, 2015), giving detailed information about individual items, including the probability that a participant will endorse a given item in a future iteration of the questionnaire. Winsteps software (Linacre, 2020a) provides a Wright map, which places items and persons on the same logit scale, showing how items perform in terms of how easily they can be endorsed by participants, and allows researchers to identify floor and ceiling effects, and also items of the same difficulty, which may indicate redundancy in measurement (Knoch & McNamara, 2015; Wilson, 2005). A person with a logit "ability" of 1.00 has a 50% chance of positively endorsing an item with a logit difficulty of 1.00. Individual item fit statistics are also

provided by Winsteps (Linacre, 2020a); these show items that may be misfitting the model, implied to be poorly functioning items.

A Rasch principal components analysis of residuals (PCAR) can be used to assess how well constructs in questionnaires conform to this unidimensional model (Raîche, 2005; Smith & Miao, 1994). The PCAR extracts the linear Rasch measure from the data and then identifies common variance in the residuals that remain. Common variance in the residuals is expressed in eigenvalues and may be an indicator of other dimensions in the questionnaire (Raîche, 2005; Smith & Miao, 1994). Rasch analyses also offer two further indices of reliability that are helpful in judging the quality of measures. These statistical indices, *reliability* and *separation*, are available for both the persons and items (Wolfe & Smith, 2007b; Wright, 1996b). *Person reliability* is equivalent to Cronbach's alpha. *Person separation* indicates the ability of the items to differentiate between different levels of the construct in the target population. *Item reliability* indicates the likelihood of the same hierarchy of items being generated from a similar population of respondents. *Item separation* is used to verify the hierarchy of items in regard to difficulty, in which low separation indicates that within a given population there is not enough variance in a construct to develop a reliable hierarchy of items. A further benefit of Rasch measurement is that it allows the construction of linear data from qualitative observations that contain a quantitative structure (Salzberger, 2010), meaning that the data can be meaningfully used for subsequent statistical procedures. In contrast, as stated above, most other statistical methods assume that typical Likert response format data are on an interval scale, and that all the items are equally adding to the measurement of the latent construct, but there is little evidence that this is in fact the case. Finally, the Rasch model provides detailed information about the functioning of each individual item, showing the relative difficulty of the item and also how well it fits into the model for measurement (Apple & Neff, 2012).

A final issue with many studies that seek to model motivation using SDT is the lack of an outcome measure, or an overreliance on a single method of measurement, a common problem in language motivation research (Al-Hoorie et al., 2021). Positive correlations may in fact be the result of common method variance (Sharma, Yetton, & Crawford, 2009). Al-Hoorie et al. (2021) argue that a vast majority of quantitative studies on motivation attempt to create models using only self-reported data, causing concerns over reliability. Within the field of L2 acquisition, there are increasing calls to employ behavioral measures as outcome variables (Oga-Baldwin et al., 2017; Yun, Hiver, & Al-Hoorie, 2018), and studies are needed that show how self-reported

motivation using both the Basic Psychological Needs Satisfaction and Frustration Scale (BPNSFS) and the Language Learning Orientation Scale (LLOS) relate to language learning outcomes. Two recent studies have considered the relationship between motivation and performance on English proficiency tests, focusing on the receptive skills of listening and reading (Apple et al., 2020; Yashima et al., 2017). Both used the L2MSS model, and Yashima et al. (2017) directly related the model to SDT, as described previously. Yashima et al. (2017) found that both internalized and external forms of motivation significantly predicted English proficiency, although internalized forms were stronger predictors in their model. Apple et al. (2020) results showed that external motivation (the ought-to self) was related to lower scores on the proficiency test, whereas internal motivation (the ideal self) was associated with higher proficiency. See also a study among South Koreans by Hiver and Al-Hoorie (2020) that linked motivation data with achievement data as an outcome.

In the current study we use the Test of English for International Communication Listening and Reading components (TOEIC L&R ETS, 2008) as an outcome variable. Although relatively unknown outside of Asia, the test is used extensively in Japan as a measure of English proficiency and is marketed as a test of business English proficiency that can be used for the advancement of one's career. It is 2 hr in duration, with approximately 45 min of listening and 75 min of reading, and a maximum scaled score of 495 for each section. ETS (2013) reported internal consistency of the listening and reading combined as .90 (Kuder–Richardson reliability index-20), with the standard errors of measurement at $\pm$ 25 points. (The Kuder Richardson reliability index is the appropriate test of reliability for multiple choice format tests.) All responses are multiple-choice, and therefore the test could be considered to be assessing comprehension of spoken and written text. Zhang (2006) investigated the reliability and generalizability of the test and claimed that it performed well on both counts. Although both the content and construct validity have been questioned (Nicholson, 2015), Powers and Powers (2015) were able to show strong correlations between performance on the TOEIC and students' self-assessment of their own ability in related, real-world tasks (perceived self-efficacy).

The Present Study

SDT has the potential to help researchers and teachers understand and enhance students' motivation in the language classroom, but accurate measurement is important. To date, to the best of our knowledge, no study has employed Rasch analysis to determine the quality of questionnaires employed in SDT research in L2 learning, despite the benefits of the Rasch model outlined above. In

addition, there has been a lack of research attempting to link these measures with performance-based outcomes such as language proficiency that could be argued to indicate levels of motivation (Al-Hoorie et al., 2021). In this study we aimed to address these gaps by conducting a Rasch analysis of the BPNSFS and the LLOS. Logit scores (scaled units when raw scores are changed to log odds ratios) from the Rasch analysis are used in regression models to assess the relationship between self-reported motivation and the TOEIC listening and reading (L&R) sections. In the context for the current study, where there is little exposure to English outside of the classroom, we hypothesized that high levels of motivation should lead to greater time spent engaged in reading and listening to English, which could be expected to manifest itself in stronger performance on the TOEIC L&R components. The following research questions drove the study:

1. To what extent do the motivational constructs measured by the BPNSFS and the LLOS fit the Rasch model for measurement?
2. How do the motivational constructs measured by the BPNSFS and the LLOS relate to performance on the TOEIC L&R components?

We are aware that the motivational measures investigated in the current study were designed for classical statistical validation using factor analysis, and not for the Rasch model. Nevertheless, following Apple and Neff (2012), we will use Rasch analysis to provide additional information regarding the relative difficulty of endorsing the different items and any issues with individual item functioning.

Method

Participants

The data for the current study were collected in a tertiary educational context in western Japan. Of a total of 878 second-year students, 600 agreed to participate in the research study. They were majoring in economics and had experienced at least six years of compulsory English education. They had also completed an additional 4.5 hr of weekly English classes in their first year of university. Students were enrolled in compulsory reading, listening, and speaking classes (each 90 min in length) that each met weekly. Each semester was comprised of 15 weeks. At the end of the second semester, students sat the TOEIC L&R and completed two motivation questionnaires (the BPNSFS and the LLOS).

Measures

The Japanese version of the BPNSFS was used for this study (Nishimura & Suzuki, 2016). As recommended by McEown and Oga-Baldwin (2019), several adaptations were made before it was administered to students. First, negatively worded items were either removed or reworded. Research has consistently shown that negatively worded items cause unwanted noise in measurement and can be responsible for up to 30% of the variance in a measure (Alexandrov, 2010). Wolfe and Smith (2007a) argue that negatively worded items create an additional dimension in the data that may represent a different construct to the one that is being targeted. Items from Agawa and Takeuchi (2016) were added to ensure that there were sufficient items for each of the three SDT constructs. Costello and Osborne (2005) suggest a minimum of three items per factor, but the inclusion of more items than necessary allows for possible removal of poorly performing ones. In addition, higher numbers of items increase reliability and increase the likelihood of replicability (Velicer & Fava, 1998). The Japanese translation of all items was checked by a native speaker of Japanese, and slight modifications were made.

As the LLOS (Noels et al., 2000) was originally designed for Canadian students of French, items were reworded for the Japanese context, translated into Japanese, and then checked by a native speaker of Japanese. As items for amotivation were not included in the original questionnaire by Noels et al. (2000), five items to measure amotivation were taken from Agawa and Takeuchi (2016). The final versions of the BPNSFS and the LLOS (Lemming & Harris, 2021a) are provided in English and Japanese in Appendices S1 and S2 in the Supporting Information online and on IRIS at https://www.iris-database.org. For both questionnaires, the Japanese version was administered to participants. Students were asked to complete the two questionnaires during their speaking class and were assured that participation was completely voluntary. They were also assured that their names and personal information would be protected.

As part of their course, the students were all required to take the TOEIC L&R, and this was used as the outcome variable. This test was administered to all the students in the department, and the scores were reflected in their class grades; therefore, they were expected to be motivated to perform well on the test. Complete data for 539 students with both motivational measures and TOEIC L&R scores were used for the regression analyses.

Analyses

The analyses described in the current study follow the procedure for validation outlined by Apple (2013) and by Apple and Neff (2012). Rasch analysis was

conducted using Winsteps software (Linacre, 2020a) on the BPNSFS and the LLOS. The procedure adopted was the same for both measures. First, a separate PCAR analysis was conducted on the items hypothesized to measure each proposed construct making up the instrument. There are several benchmarks provided in the literature for PCAR. Smith and Miao (1994) argue that unexplained variance in the first contrast with eigenvalues greater than 1.4 may indicate the presence of other dimensions, although Wright (1996a) argues that values up to 2.0 are acceptable for unidimensional measures. The variance explained by the measures should be greater than 50% (Linacre, 2020b). Eigenvalues are the same as those used in factor analysis, but the interpretation is different. In factor analysis, high eigenvalues show shared variance by items and suggest that they are measuring the same construct, whereas in Rasch PCAR, the eigenvalues are of the residuals, and therefore higher values suggest correlation among residuals, or the presence of a second dimension in the data.

Following PCAR, the reliability and separation statistics for each intended construct were considered. Reliability statistics can be interpreted in the same way as Cronbach's alpha, with values above .80 generally considered acceptable (Field, 2009). Linacre (2020b) claims that person separation should be greater than 2 to distinguish high and low performers, and that item separation should be greater than 3, although Apple and Neff (2012) argue that item separation values greater than 5 are preferable to show different levels of a construct within a given population. The next analysis considered the effectiveness of the 6-point Likert categories. Put simply, this analysis examines whether all the points on the scale are used, and whether there are real differences between levels as indicated by the scale. Wolfe and Smith (2007b) recommend that for an acceptable scale there should be: at least 10 observations per Likert category across the data set; average person measures for each category should be higher than the mean score of the level below; outfit mean squares should be lower than 2.0; and gaps in difficulty between levels should be greater than 0.59 and less than 5 logits.

The final analysis investigated the fit of individual items to the Rasch model. Misfitting items generally suggest that they are not effectively measuring the target construct (Waugh & Chapman, 2005; Wolfe & Smith, 2007a). Following Wolfe and Smith (2007b) and Smith (2000), mean square fit (MNSQ) values of between 0.6 and 1.4, and standardized *z*-scores (ZSTD) values between −2 and +2 were considered acceptable. Ideal mean square fit values are 1.00 and ideal *z*-scores are zero. High mean square values (greater than 1) indicate erratic and unpredictable response patterns that are

underfitting the model, whereas low values (less than 1) highlight responses that are too predictable and are therefore overfitting the model. Infit values are weighted to give information about the respondents close to their level of endorsability for a given item, whereas outfit values are unweighted and provide information about respondents who are either far above or below the threshold for the given item. Therefore, high infit values are of more concern than high outfit values with regard to quality of measurement because they indicate poorly written items (Apple & Neff, 2012). Bond et al. (2021) explain the difficulties in interpreting fit values. When sample sizes are low, it is more likely that the mean square fit statistics will show misfit to the model, but although increases in sample size benefit the infit mean square, they negatively influence the standardized *z*-scores. Based on their moderate sample size of approximately 200 participants, Apple and Neff (2012) used standardized values, but with 600 participants in the current study, the decision was more difficult. Wolfe and Smith (2007b) suggest that with fewer than 1,000 respondents, *z*-score values greater than 2.00 indicate misfitting items. We decided to highlight items that did not fit the criteria for both standardized *z*-scores and mean square values.

To answer the first research question, regarding the validity of the SDT measures, various statistics are presented and discussed in relation to the benchmark values presented above. It should be noted that, as with all statistical analyses, these benchmarks are guidelines rather than hard and fast rules, and the fact that a measure meets the suggested criteria does not guarantee its validity. Although these benchmarks are widely used in the Rasch literature, Bond et al. (2021) stress that careful consideration is needed rather than simply adhering to cutoff statistics, and Linacre (2020b) also states that these benchmarks should be used for guidance, to help researchers make informed decisions about individual items in questionnaires.

Results

The data for the current study (Leeming & Harris, 2021b) have been made available online at IRIS (https://www.iris-database.org) and OSF https://mfr.osf.io/render?url=https%3A%2F%2Fosf.io%2Ffs86b%2Fdownload.

Rasch Analysis of Self-Determination Theory Measures

The first research question concerned how well two scales used in SDT research would perform in a Rasch analysis. The results for the two scales, the BPNSFS and the LLOS, are presented in the following subsections.

Table 1 Results of principal components analysis of residuals for the Basic Psychological Needs Satisfaction and Frustration Scale

Construct	% variance explained	% unexplained variance	% first contrast explained variance	First contrast eigenvalue
Autonomy	59.0	41.0	11.4	1.67
Competence	62.4	37.6	9.6	1.52
Relatedness	66.6	33.4	10.7	1.83

Table 2 Person and item separation and reliability for the constructs of the Basic Psychological Needs Satisfaction and Frustration Scale

Construct	Person separation	Person reliability	Item separation	Item reliability
Autonomy	2.29	.84	9.16	.99
Competence	2.49	.86	9.37	.99
Relatedness	2.87	.89	6.38	.98

The Basic Psychological Needs Satisfaction and Frustration Scale

For the BPNSFS, a separate Rasch PCAR was run on each of the three hypothesized constructs: autonomy, competence, and relatedness. All three showed strong unidimensionality, with variance explained ranging from 59% to 67% and low eigenvalues for the first contrast (Table 1).

We then examined the person and item separation and reliability for each construct (see Table 2). Person separation and reliability were strong for each of the three constructs, with reliability above .80 and separation values above the benchmark of 2.0 recommended by Linacre (2020b). This means that the questionnaire was effective in highlighting different levels of the constructs among participants. The item reliabilities were also strong with values of .98 and above, indicating consistency among respondents in answering the questions. Item separation was strong, with values ranging from 6.38 to 9.37, above the desired value of 5.0 for all three constructs (Apple & Neff, 2012), suggesting that the items effectively measure different levels of the constructs within the target population.

Wright maps were examined for the three constructs (Figures 2–4). Persons are represented on the left by the symbols "#" (representing five respondents) and "." (representing one to four respondents). Those respondents with more of the construct (e.g., greater feelings of autonomy in Figure 2) are at the top. Items are represented by their number and also the Likert category choice of 1

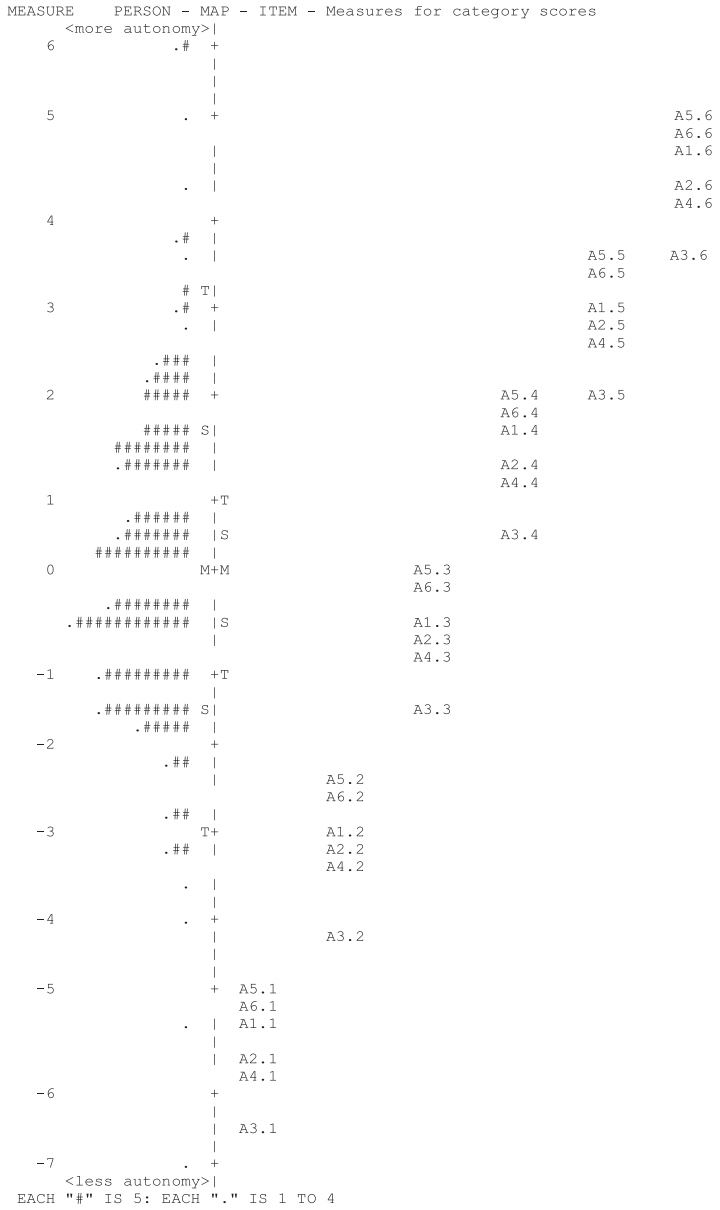

Figure 2 Wright map for autonomy.

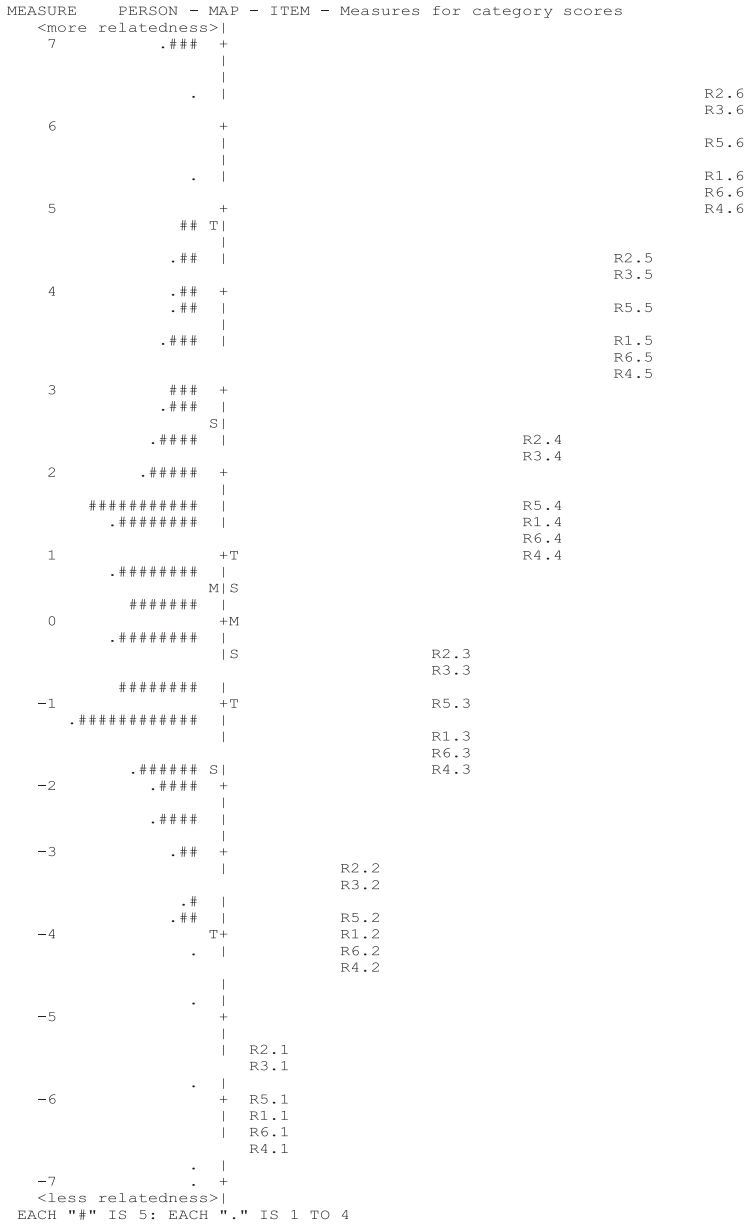

Figure 3 Wright map for relatedness.

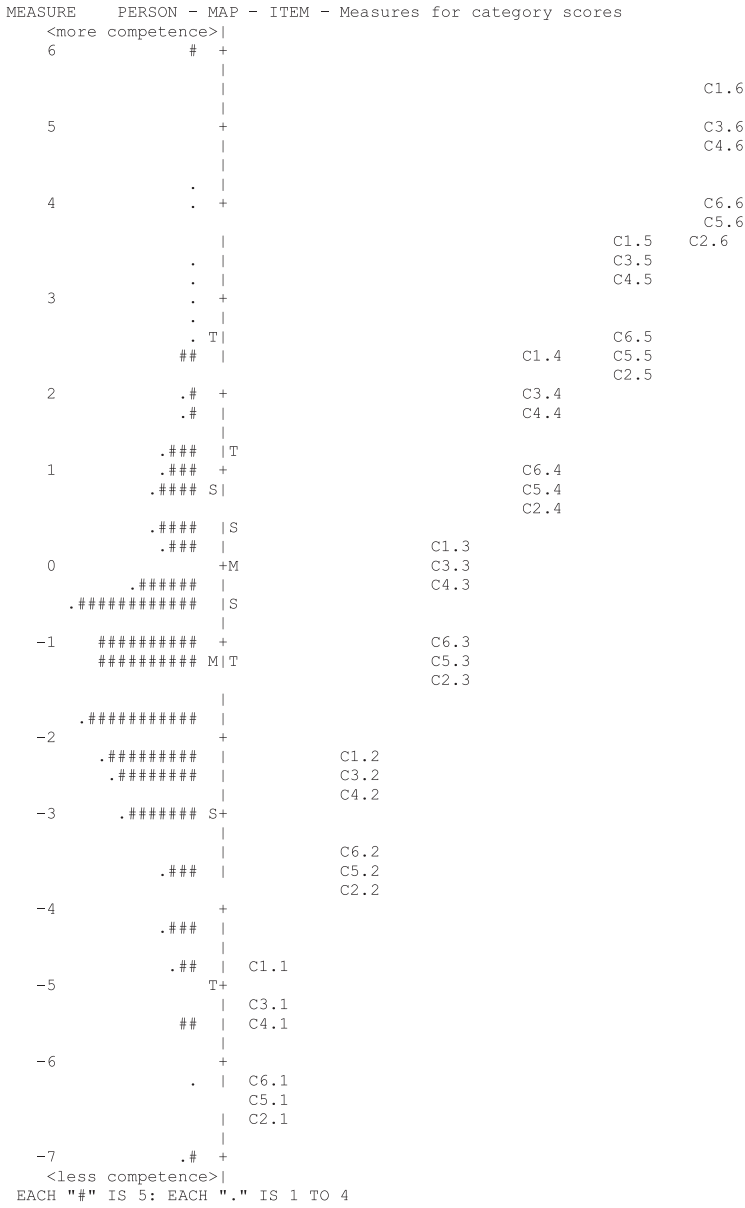

Figure 4 Wright map for competence.

to 6. Positioning of items higher up the map indicates that they are harder for persons to positively endorse, and therefore agreement is indicative of more (relative to items lower down on the map) of the latent construct. For example, A3.4 in Figure 1 indicates item 3 for autonomy, and the *slightly agree* choice category. When a person and item have the same logit score (are at the same level on the map), there is a 50% probability of the person endorsing the item at that particular point on the Likert scale (Wilson, 2005). Therefore, using A3.4 as an example, a person at the same level on the map as this item has a 50% probability of endorsing the *slightly agree* category.

The autonomy items (Figure 2) show a reasonable spread, with A3 being the easiest to endorse and A5 the most difficult; the item and person means are the same, suggesting that items are well matched to these respondents. There is a degree of redundancy, however, with some overlap between A3 and A5 at the higher end of the figure. Figure 2 shows that a rating of *agree* (5) on A3 is equivalent to a rating of *slightly agree* (4) on A4. This difference in interpretation of raw scores highlights the dangers of simply adding Likert category response scores in order to measure a construct.

Figure 3 shows the Wright map for relatedness items, showing that they were generally easier for the participants to endorse, with the mean for persons above the mean for items. R4 was the easiest to endorse, and R2 was the most difficult. A large number of people appear above the most difficult items, indicating a ceiling effect.

The Wright map for the competence items is shown in Figure 4. Although the items show reasonable spread, again some items overlap, with C1 *slightly agree* (4) harder to endorse than C2 *agree* (5), and of equal difficulty with C5 *agree* (5). Generally, C2 is the easiest for participants to endorse and C4 the hardest, although the low mean for the participants suggests that overall, the items were difficult for these students to endorse.

The next analysis concerned the functioning of the 6-point Likert rating scale and investigated whether respondents used the entire scale, and whether difficulty increased incrementally with the response category chosen. The Andrich threshold is the value at which there is a 50% chance of selecting that category or the one below (e.g., *disagree* or *strongly disagree*). The category measure indicates where that response is on the logit scale (see the Wright maps in Figures 2–4). Based on Wolfe and Smith's (2007b) criteria, the results suggest that the six Likert categories are functioning well for the items in each of the constructs that make up the BPNSFS. See Appendix S3 in the Supporting Information online for the table of results.

Table 3 Rasch item statistics for the Basic Psychological Needs Satisfaction and Frustration Scale

Item	Measure	*SE*	Infit *MNSQ*	Infit *ZSTD*	Outfit *MNSQ*	Outfit *ZSTD*	PMC
Autonomy							
A5	.65	.06	0.96	−0.66	0.93	−1.11	.75
A6	.55	.06	1.00	−0.02	0.94	−0.95	.78
A1	**.20**	**.06**	**1.28**	**4.24**	**1.24**	**3.73**	**.68**
A2	**−.18**	**.06**	**0.79**	**−3.82**	**0.80**	**−3.54**	**.82**
A4	−.21	.06	0.99	−0.18	1.00	0.00	.74
A3	−1.01	.06	0.94	−1.07	0.97	−0.55	.79
Competence							
C1	.91	.07	1.02	0.42	0.99	−0.23	.80
C3	.51	.06	1.10	1.68	1.04	0.64	.81
C4	.31	.06	0.94	−0.98	0.91	−1.54	.81
C6	−.44	.06	1.01	0.18	1.02	0.39	.77
C5	−.62	.06	0.90	−1.65	0.93	−1.27	.79
C2	−.66	.06	1.10	1.60	1.09	1.46	.76
Relatedness							
R2	.60	.07	1.13	2.12	1.12	1.89	.83
R3	.57	.07	0.91	−1.53	0.91	−1.46	.84
R5	**.02**	**.07**	**0.86**	**−2.47**	**0.84**	**−2.83**	**.86**
R1	**−.27**	**.07**	**0.85**	**−2.72**	**0.85**	**−2.51**	**.85**
R6	−.42	.07	0.91	−1.60	0.91	−1.48	.84
R4	**−.49**	**.07**	**1.27**	**4.19**	**1.29**	**4.46**	**.81**

Note. Bold font indicates misfitting items. *MNSQ* = mean square; ZSTD = standardized z-scores; PMC = Point measure correlation.

The final area of interest is the functioning of individual items in relation to the Rasch model of measurement, with misfitting items suggestive of potential measurement problems. Table 3 shows the item difficulty, or how easily an item can be endorsed by participants (Measure column), and fit statistics for the 18 items that form the BPNSFS. Point measure correlations highlight polarity of items and show whether the items are working in the direction expected. Thus, negative figures or those close to zero suggest that an item is not performing well (Bond et al., 2021). According to Linacre (2006), any figure greater than +.4 suggests that indicator polarity is consistent in the scale. Items that do not meet the benchmarks for either mean square (*MNSQ*) or standardized values (*ZSTD*) are shown in bold. Although the competence items show good fit to

Table 4 Results of principal components analysis of residuals for the Language Learning Orientation Scale

Motivation construct	% variance explained	% unexplained variance	% first contrast explained variance	First contrast eigenvalue
Intrinsic	66.5	33.5	8.4	1.75
Integrated	62.2	37.8	11.6	1.53
Identified	61.5	38.5	10.9	1.42
Introjected	54.6	45.4	16.6	1.83
External	45.9	54.1	23.8	2.20
Amotivation	70.0	30.0	10.6	1.77

the model, the autonomy and relatedness constructs each have several items of concern. Possible reasons for misfit are presented in the discussion section.

To determine the influence of misfitting items on measurement, the disattenuated correlations for each construct measured with and without the misfitting items were correlated. A bivariate correlation analysis was performed, and the correlations were significant ($p < .001$) for both autonomy ($r = .97$, 95% CI [.96, .97]) and relatedness ($r = .96$, 95% CI [96, 96]), indicating that removal of the items does little to change the results for persons.

The Language Learning Orientation Scale

The procedure described for the BPNSFS was repeated for the LLOS. A Rasch PCAR was conducted for each of the six constructs hypothesized to exist in the instrument: *intrinsic motivation*, *integrated motivation*, *identified motivation*, *introjected motivation*, *external motivation*, and *amotivation*. The results in Table 4 show that five of the constructs show reasonable unidimensionality, with variance explained ranging from 55% to 70% and low eigenvalues for the first contrast. The weakest construct was external motivation, where the variance explained fails to meet the threshold value of 50% (Linacre, 2020b), and the first eigenvalue was greater than 2, suggesting that it is not unidimensional (Wright, 1996a).

Following the PCAR, we examined the person and item separation and reliability for each construct. Person separation and reliability were strong for four of the six constructs (Table 5), with reliability above .80 and separation above the recommended benchmark of 2.00. Introjected and external motivation showed weak reliability (below .80), and also poor person separation. This suggests that respondents were inconsistent in their responses to these items, and also that the items were not effective in separating the students

Table 5 Person and item separation and reliability for the constructs of the Language Learning Orientation Scale

Motivation construct	Person separation	Person reliability	Item separation	Item reliability
Intrinsic	3.04	.90	6.84	.98
Integrated	2.27	.84	5.48	.97
Identified	2.24	.83	10.88	.99
Introjected	1.92	.79	6.90	.98
External	1.47	.68	8.99	.99
Amotivation	2.49	.86	4.80	.96

into different groups. Both item reliability and item separation were strong for five of the six constructs, with values for the latter ranging from 5.48 to 10.88.

Wright maps were examined for the six constructs of the LLOS (see Figures 5–10). The intrinsic motivation items showed good spread around the mean, although the mean for persons was slightly lower than that for items. INT4 was the easiest to endorse, and INT5 was the most difficult. The integrated motivation items also showed reasonable spread, although the mean for persons was a whole logit lower than the mean for items, suggesting that these items were difficult for participants to endorse. ITE5 was the easiest to endorse, and ITE4 the most difficult. Although the identified motivation items also showed reasonable spread and were well matched to the participants, there were some issues with overlapping items. IDE5 was considerably more difficult to endorse than IDE1, and a rating of *agree* for IDE5 was harder to endorse than a rating of *strongly agree* for IDE1. The introjected motivation items showed reasonable spread around the mean, and again were well matched to the respondents. ITR4 was the hardest to endorse, and ITR1 was the easiest. The external motivation items showed a good spread and were generally easy for participants to endorse, with the mean for persons slightly higher than the mean for items. There were some issues with difficulty of individual items, particularly in the middle of the map. For EXT2, the most difficult item, *slightly disagree* was harder to endorse than *slightly agree* for EXT1 and EXT3. The Wright map for the amotivation items revealed a serious floor effect, with a large number of students showing very low levels of amotivation below the minimum item, suggesting either that these items were not effective in measuring the construct within these participants, or that amotivation was not present in the majority of participants. The mean for persons was approximately 3 logits below the mean for items, suggesting that these items were very difficult for the participants to

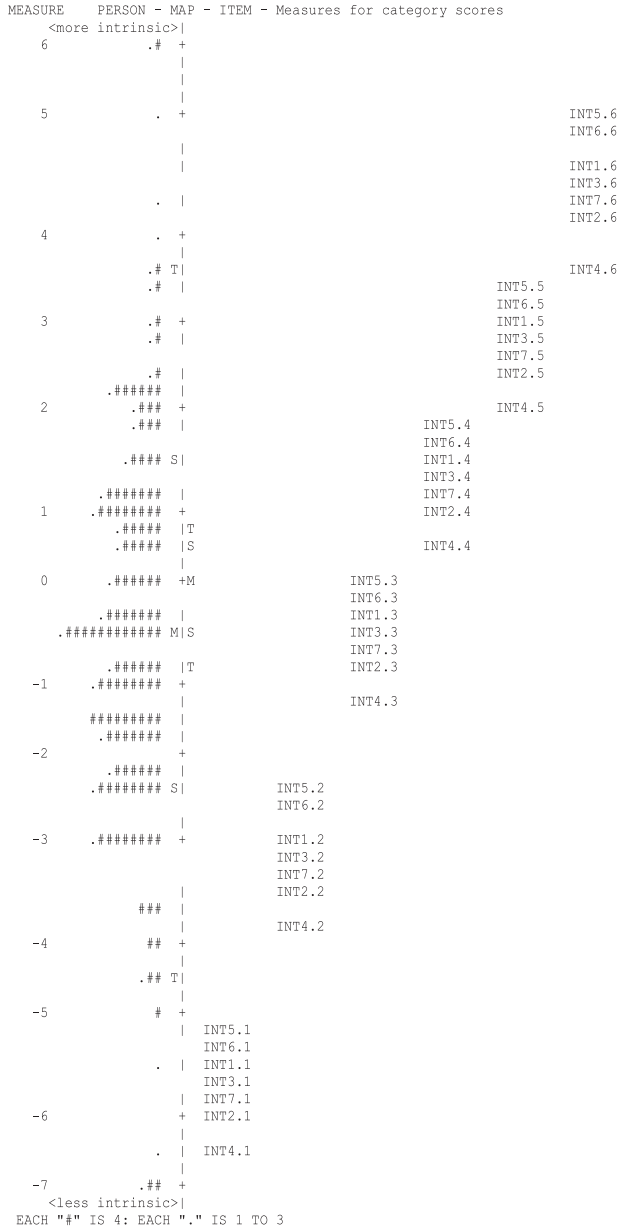

Figure 5 Wright map for intrinsic motivation.

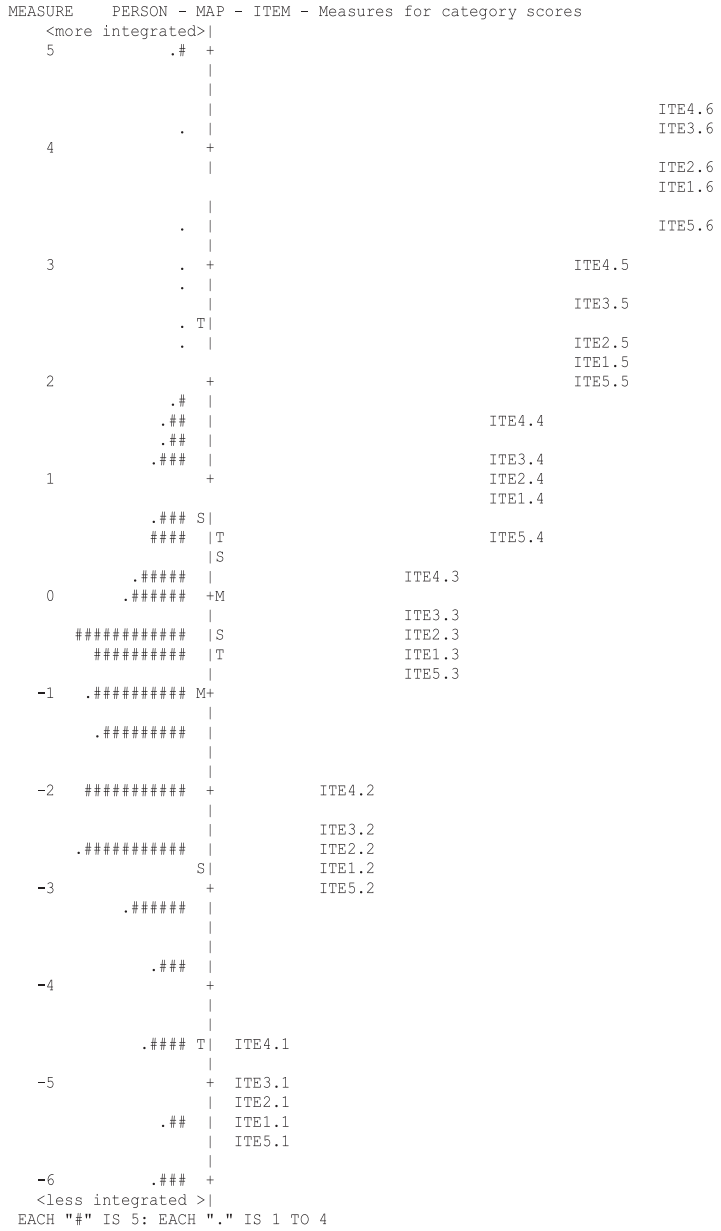

Figure 6 Wright map for integrated motivation.

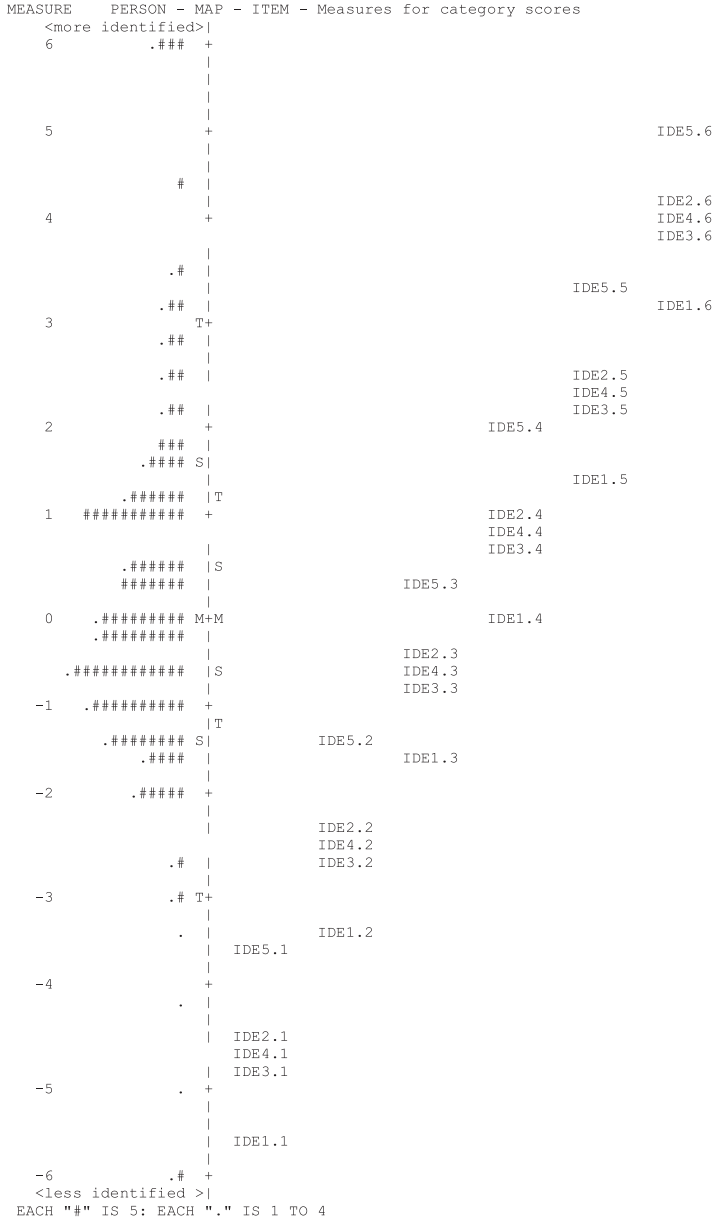

Figure 7 Wright map for identified motivation.

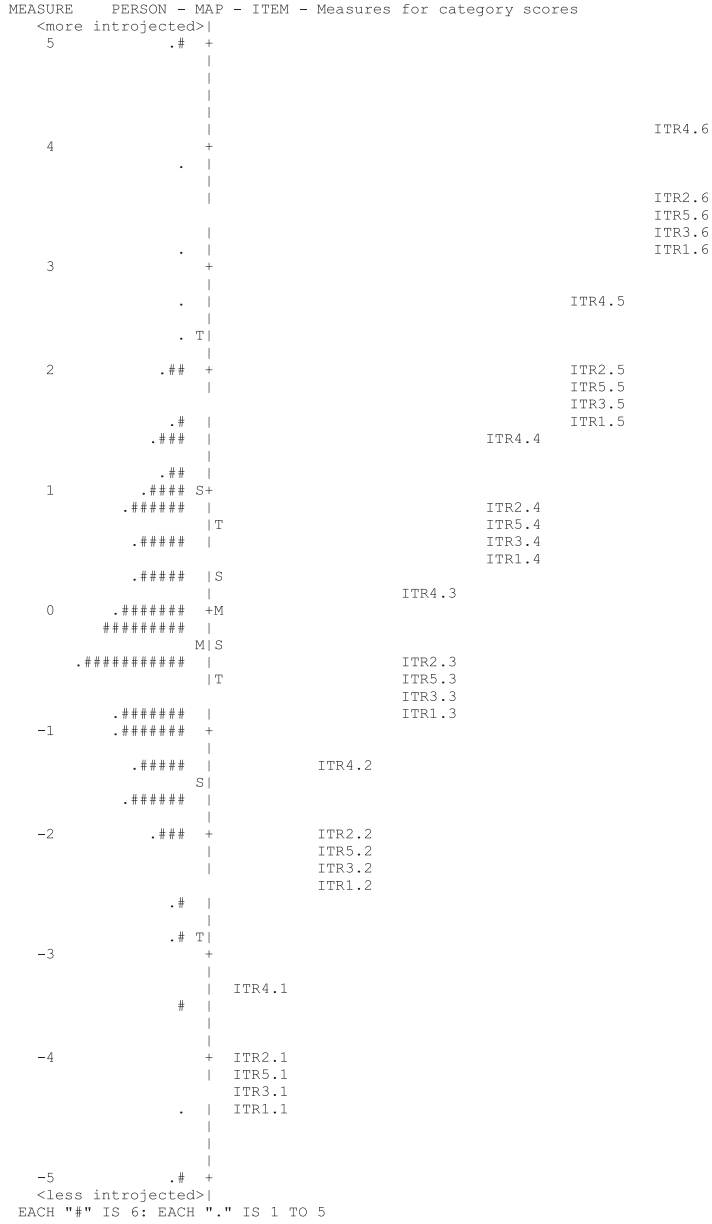

Figure 8 Wright map for introjected motivation.

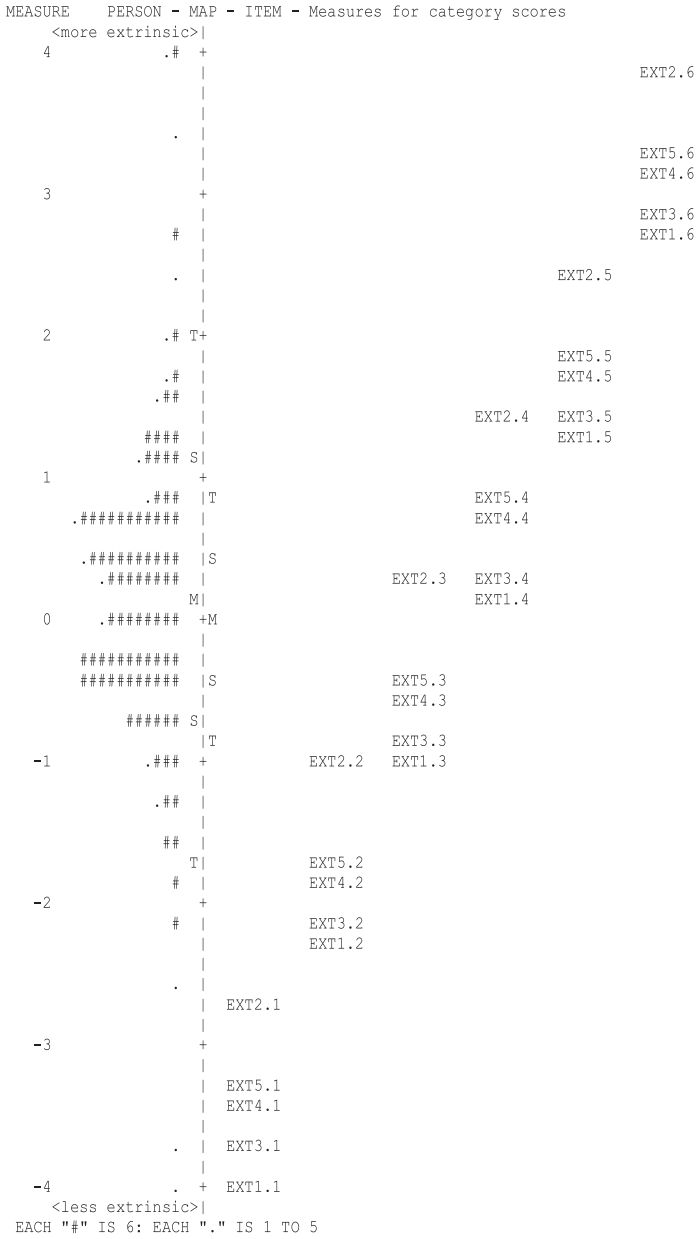

Figure 9 Wright map for external motivation.

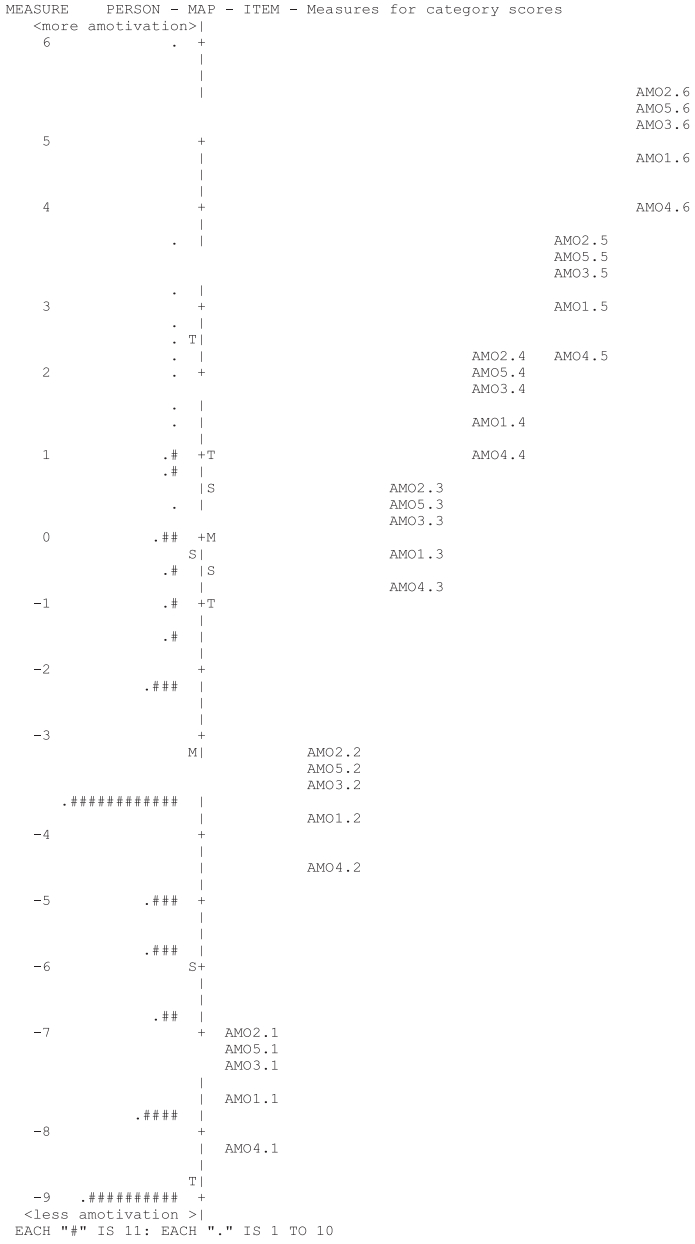

Figure 10 Wright map for amotivation.

endorse. AMO4 was the easiest to endorse, and AMO2 was the most difficult; there was also some overlap between different category responses for these items.

Appendix S4 in the Supporting Information online shows the results of the category functioning for the LLOS. Generally, the results fit the criteria previously outlined, although the amotivation construct was heavily weighted toward negative responses, with only 1% of respondents using the *strongly agree* (6) category, and 35% of all responses using the *strongly disagree* (1) rating. This suggests that respondents found it hard to agree with most of the statements in this construct, but this would be expected from a measure of amotivation, in which a complete absence of motivation is being measured.

Finally, individual items were analyzed for fit to the Rasch model. Table 6 presents the results, with bold indicating poorly fitting items based on the criteria outlined above. The table shows that there are a number of potentially problematic items in each of the constructs. As with the BPNSFS, possible reasons for misfit are presented in the discussion section.

To determine the influence of misfitting items on measurement, the disattenuated correlations for each construct were measured both with and without the misfitting items. A bivariate correlation analysis was performed, and all the correlations were significant ($p < .001$) (intrinsic, $r = .98$, 95% CI [.97, .98]; integrated, $r = .95$, 95% CI [.94, .96]; identified, $r = .89$, 95% CI [.87, .90]; introjected, $r = .92$, 95% CI [.91, .93]; external, $r = .75$, 95% CI [.71, .78]; amotivation, $r = .95$, 95% CI [.94, .96]). The comparatively weak correlation for external motivation is further evidence that the removed items were perhaps measuring a different construct, and the slightly numerically lower correlation coefficients for identified and introjected motivation could also suggest weakness in some of the items measuring these constructs.

Self-Determination Theory and TOEIC Listening and Reading

The second research question concerned the relationship between the measures of motivation and performance on the TOEIC L&R. The latter provides an outcome variable absent from many previous studies of language motivation (as noted by Al-Hoorie et al., 2021). The logit scores derived from the Rasch analyses of the measures of motivation outlined above were added to regression models with, separately, listening scores and reading scores from the tests ($N = 539$). The independent variables of motivation were based on the revised measures, with the misfitting items revealed by the Rasch analyses, highlighted in Tables 3 and 6, removed prior to the correlation and regression analyses.

Table 6 Rasch item statistics for items on the Language Learning Orientation Scale

Item	Measure	*SE*	Infit *MNSQ*	Infit *ZSTD*	Outfit *MNSQ*	Outfit *ZSTD*	PMC
Intrinsic motivation							
5	**.49**	**.06**	**1.27**	**4.06**	**1.21**	**3.25**	**.79**
6	.48	.06	1.10	1.57	1.05	0.88	.81
1	.13	.06	1.06	1.05	1.04	0.69	.80
3	**.06**	**.06**	**0.87**	**−2.29**	**0.85**	**−2.65**	**.84**
7	**−.06**	**.06**	**0.88**	**−2.01**	**0.86**	**−2.37**	**.83**
2	−.27	.06	0.89	−1.88	0.90	−1.68	.82
4	−.83	.06	0.99	−0.19	1.02	0.41	.81
Integrated motivation							
4	.56	.06	1.17	2.59	1.11	1.80	.77
3	.17	.06	0.96	−0.56	0.95	−0.79	.79
2	**−.10**	**.06**	**0.79**	**−3.66**	**0.77**	**−4.07**	**.83**
1	−.19	.06	0.88	−2.01	0.92	−1.26	.80
5	**−.43**	**.06**	**1.21**	**3.25**	**1.21**	**3.31**	**.77**
Identified motivation							
5	**.99**	**.06**	**1.33**	**5.04**	**1.26**	**4.10**	**.72**
2	.08	.05	1.13	2.11	1.11	1.85	.81
4	**.01**	**.05**	**0.85**	**−2.65**	**0.84**	**−2.86**	**.82**
3	**−.14**	**.05**	**0.77**	**−4.20**	**0.77**	**−4.22**	**.82**
1	−.95	.05	0.94	−1.10	0.92	−1.33	.76
Introjected motivation							
4	.67	.05	1.13	2.21	1.06	1.09	.72
2	**.01**	**.05**	**0.90**	**−1.86**	**0.87**	**−2.26**	**.77**
5	−.14	.05	1.10	1.77	1.12	1.97	.73
3	**−.22**	**.05**	**0.87**	**−2.33**	**0.87**	**−2.30**	**.77**
1	−.33	.05	0.99	−0.13	1.03	0.56	.71
External motivation							
2	**.73**	**.05**	**1.19**	**3.13**	**1.18**	**2.94**	**.63**
5	.12	.04	0.96	−0.76	0.95	−0.93	.70
4	**−.04**	**.04**	**0.68**	**−6.31**	**0.67**	**−6.39**	**.74**
3	−.33	.04	0.94	−1.15	0.93	−1.23	.67
1	**−.48**	**.04**	**1.22**	**3.66**	**1.23**	**3.69**	**.57**
Amotivation							
2	**.44**	**.09**	**0.88**	**−1.52**	**0.71**	**−3.68**	**.85**
5	.25	.09	1.12	1.47	1.16	1.76	.85
3	**.24**	**.09**	**0.78**	**−2.96**	**0.65**	**−4.62**	**.88**
1	−.14	.09	0.91	−1.19	0.86	−1.71	.84
4	**−.80**	**.08**	**1.30**	**3.67**	**1.42**	**4.73**	**.86**

Note. Bold font indicates misfitting items. *MNSQ* = mean square; ZSTD = standardized z-scores; PMC = Point measure correlation.

Table 7 Multiple regression model predicting the Listening component of the Test of English for International Communication with the Basic Psychological Needs Satisfaction and Frustration Scale

	b	*SE*	b^*	*t*	*p*	95% CI	*VIF*
Intercept	258.62	4.67	.00	55.35	< .001	[249.44, 276.79]	
Aut.	9.08	2.64	.22	3.44	.001	[3.89, 14.27]	2.37
Rel.	−2.64	1.41	−.10	−1.87	.06	[−5.42, 0.13]	2.21
Comp.	3.49	2.50	.09	1.40	.16	[−1.42, 8.41]	1.70

Note. b = unstandardized regression coefficient; b^* = standardized regression coefficient; *VIF* = variance inflation factor; Aut. = autonomy; Rel. = relatedness; Comp. = competence.

Descriptive statistics from the TOEIC L&R and for the nine variables measuring motivation are presented in Appendices S5 and S6, respectively, of the Supporting Information online. Examination of the histograms for TOEIC L&R scores showed normal distributions, although the reading scores showed slight positive skew. Field (2009) argues that significant skew and kurtosis are likely with large sample sizes, and recommends consideration of histograms and absolute values. On this basis, the TOEIC L&R data can be considered to have a normal distribution. The correlation matrix for all the variables in the study can be seen in Appendix S7 online. There is continued debate over correlation values and collinearity (Dormann et al., 2013), with suggestions that values greater than .70 or greater than .90 (Tabachnick & Fidell, 2007) indicate problems with the variables. Therefore, a moderate value of .80 was selected for this study, and on this basis, correlations were deemed acceptable.

Data were screened following the procedure for regression analysis outlined by Field (2009), and univariate and multivariate outliers were removed from the data (a total of 37 participants). A multiple linear regression model was constructed with scores from the TOEIC listening section as the dependent variable (DV) and the three constructs of autonomy, relatedness, and competence as predictor variables (Table 7). As no a priori hypothesis existed regarding the relationship between independent variables (IVs) and the DV, the entry method was used for IVs (see Field, 2009). Apha values below .05 were used as the benchmark for significance. The model accounted for 6% (R^2; adjusted $R^2 = .05$) of the variance in the listening score, $F(3, 535) = 10.82$, $p < .001$, with the observed effect being large enough to meet the threshold previously determined by sensitivity power analysis ($R^2 < .02$). An R^2 value of .06 is equivalent to a correlation of approximately .24 and can be classified

Table 8 Multiple regression model predicting the Reading component of the Test of English for International Communication with the Basic Psychological Needs Satisfaction and Frustration Scale

	b	*SE*	b^*	*t*	*p*	95% CI	*VIF*
Intercept	198.94	4.31	.00	46.18	< .001	[190.48, 207.40]	
Aut.	9.97	2.43	.26	4.09	< .001	[5.19, 14.75]	2.37
Rel.	−2.49	1.30	−.10	−1.91	.06	[−5.05, 0.07]	2.21
Comp.	2.71	2.31	.07	1.17	.24	[−1.83, 7.24]	1.70

Note. b = unstandardized regression coefficient; b^* = standardized regression coefficient; *VIF* = variance inflation factor; Aut. = autonomy; Rel. = relatedness; Comp. = competence.

as a small effect size (Plonsky & Oswald, 2014). Autonomy was the only significant predictor of TOEIC listening performance. Variance inflation factor values were well below 10, suggesting that multicollinearity was not an issue with these variables (Field, 2009).

A second multiple linear regression model was constructed with TOEIC reading results as the DV and the three constructs autonomy, relatedness, and competence as predictor variables (Table 8). The model accounted for 7% (R^2; adjusted $R^2 = .07$) of the variance in the listening score, $F(3, 535) = 13.51$, $p < .001$. An R^2 value of .07 is equivalent to a correlation of approximately .26 and can be classified as a small effect size (Plonsky & Oswald, 2014). Again, autonomy was the only significant predictor of TOEIC reading performance.

The regression procedure was repeated for the LLOS variables. A multiple linear regression model was constructed with TOEIC listening as the DV and the six constructs from the LLOS as predictor variables (Table 9). The model accounted for 13% (R^2; adjusted $R^2 = .12$) of the variance in the listening score, $F(6, 532) = 12.70$, $p < .001$, with the observed effect being large enough to meet the threshold previously determined by sensitivity power analysis ($R^2 < .02$). An R^2 value of .13 is equivalent to a correlation of approximately .36 and can be classified as a small to medium effect size (Plonsky & Oswald, 2014). The significant predictors of listening scores were intrinsic motivation, identified motivation, introjected motivation, and amotivation. Introjected motivation and amotivation had a negative relationship with the DV.

The fact that introjected motivation has a strong negative relationship with the outcome variable ($b^* = -.24$, 95% CI [.87, .90]) and yet near zero correlation (see Appendix S7 in the Supporting Information online) suggests that it is a suppressor variable (Pedhazur, 1997; Tabachnick & Fidell, 2007). A

Table 9 Multiple regression model predicting the Listening component of the Test of English for International Communication with the Language Learning Orientation Scale

	b	*SE*	b^*	*t*	*p*	95% CI	*VIF*
Intercept	233.62	5.61	.00	41.61	< .001	[222.38, 244.41]	
Intrinsic	6.53	2.45	.19	2.66	.01	[1.71, 11.35]	3.19
Integrat.	1.49	2.00	.05	0.74	.46	[−2.45, 5.42]	2.34
Ident.	3.49	1.59	.13	2.19	.03	[0.36, 6.61]	2.24
Introj.	−11.12	2.90	−.24	−3.84	< .001	[−16.81, −5.43]	2.33
Extr.	0.06	1.00	.00	0.06	.95	[−1.90, 2.02]	1.50
Amot.	−3.32	0.82	−.19	−4.05	< .001	[−4.93, −1.71]	1.31

Note. b = unstandardized regression coefficient; b^* = standardized regression coefficient; *VIF* = variance inflation factor; Intrinsic = intrinsic motivation; Integrat. = integrated motivation; Ident. = identified motivation; Introj. = introjected motivation; Extr. = external motivation; Amot. = amotivation.

suppressor variable improves the overall model by removing the variance that is shared by predictor variables but not associated with the criterion variable, therefore effectively removing noise from the model (Pedhazur, 1997). Tabachnick and Fidell (2007) describe it as a variable that "enhances the importance of other IVs by virtue of suppression of irrelevant variance in them" (p. 155). The implications of this will be discussed in the next section. The standardized regression coefficient values show that, with the exclusion of the introjected motivation construct, amotivation and intrinsic motivation account for the most variance in the DV, followed by identified motivation. Overall, the model accounts for a small but significant part of the variance in listening scores.

A multiple regression model was constructed with reading as the DV and the six LLOS constructs as IVs (Table 10). The model accounted for 11% (R^2; adjusted $R^2 = .10$) of the variance in the listening score, $F(6, 532) = 10.65, p < .001$. An R^2 value of .11 is equivalent to a correlation of approximately .33 and can be classified as a small to medium effect size (Plonsky & Oswald, 2014). The significant predictors were identified motivation, introjected motivation, and amotivation. As with the results for the TOEIC listening scores, introjected motivation appears to be working as a suppressor variable. In contrast to the results for the TOEIC listening scores, intrinsic motivation was not a significant predictor of performance.

Table 10 Multiple regression model predicting the Reading component of the Test of English for International Communication with the Language Learning Orientation Scale

	b	*SE*	b^*	*t*	*p*	95% CI	*VIF*
Intercept	179.11	5.26	.00	34.05	< .001	[168.77, 189.44]	
Intrinsic	2.30	2.30	.07	1.00	.32	[−2.22, 6.82]	3.19
Integrat.	2.88	1.88	.10	1.53	.13	[−0.82, 6.57]	2.34
Ident.	4.67	1.49	.19	3.13	.002	[1.74, 7.60]	2.24
Introj.	−10.40	2.72	−.24	−3.83	< .001	[−15.74, −5.07]	2.33
Extr.	1.01	0.94	.05	1.08	.28	[−0.83, 2.85]	1.50
Amot.	−2.46	0.77	−.15	−3.20	< .001	[−3.97, −0.95]	1.31

Note. b = unstandardized regression coefficient; b^* = standardized regression coefficient; *VIF* = variance inflation factor; Intrinsic = intrinsic motivation; Integrat. = integrated motivation; Ident. = identified motivation; Introj. = introjected motivation; Extr. = external motivation; Amot. = amotivation.

Discussion

The Functioning of the Two Questionnaires: BPNSFS and LLOS

The first research question was concerned with the functioning of the BPNSFS and LLOS measures as investigated through Rasch analysis. The constructs measured by the BPNSFS showed strong unidimensionality, and the measures were effective in differentiating between different levels of autonomy, competence, and relatedness in the population for this study. The six Likert categories functioned well, although there were items for both autonomy and relatedness that appeared to misfit the Rasch model. Item A1 shows underfit and A2 shows overfit to the model (as stated above, generally, overfit is of less concern for measurement than underfit). One possible reason for A2's different behavior may lie in the fact that the wording has no reference to "choice." For example, A5 ("I feel that my decisions reflect what I really want") allows for choice. On the other hand, A2 ("I have a strong desire to learn") alludes to a fixed emotional state. Reasons why A1 ("I feel a sense of choice in the things I undertake") is misfitting the model are less clear. With regard to relatedness, R1 ("I feel comfortable working with my classmates"), R4 ("There is a friendly atmosphere in English class"), and R5 ("I get along well with my classmates in English class") also show weak fit to the model, with R1 and R5 overfitting and R4 underfitting. Unlike other items, these refer to positive feelings, which may be why they are behaving differently.

Although most of the constructs in the LLOS appeared to be unidimensional, results for external motivation suggest more than one dimension. This may be explained by the disparate nature of the items, whose topics include the importance of getting course credit and attaining high test scores, being told that it is important to study, and the perceived need to get a job or a high salary. Reliability and person separation were also weak for this construct. In addition, although the Likert categories generally worked satisfactorily, a very small number of students used the *strongly agree* option. Also, some individual items did not fit to the Rasch model.

Intrinsic motivation items 3 and 7 overfit, whereas 5 underfit. Item 3 mentions the "high" of learning English (in Japanese this was translated as *shigeki*, meaning "stimulation"), which may be hard for students in formal educational contexts to relate to. Item 7 refers to a "challenge," which was translated as *yarigai* in Japanese. The latter contains a nuance of being something that is "challenging" but also "rewarding," and therefore may have been interpreted as a difficult thing that is worthwhile, but not necessarily held to be intrinsically motivating. Item 5 mentions "hearing" English spoken, and it is possible that in this educational context, which often focuses intensively on reading and discrete grammar, students have limited experience of hearing spoken English, especially authentic spoken input. The Japanese translation includes the phrase *eigo ga hanasareteiru koto no wo kiku* ("to listen to English being spoken"). In Japanese *kiku* can mean both "to hear" and "to listen," and therefore may have caused confusion for students, as it covers both unintentional (hearing) and intentional (listening) aspects. Although students today have far greater access to English through media such as Netflix and YouTube, they often rely on dubbed or subtitled shows. As a result, despite being exposed to and hearing English through these media, they may not necessarily listen and focus on the English language input. This may have influenced their responses to these questions.

Items 2 (overfit) and 5 (underfit) from the integrated motivation construct were also examined. These two items are unique in referring to "self-image" and "identity," which may complicate their interpretation by participants. Indeed, this may be an example of the difficulty of transposing existing instruments across cultural contexts without consideration for cultural differences. Three items for identified motivation misfit the model: items 3 and 4 (overfit), and item 5 (underfit). There appears to be a split with the items for identified motivation, with the first two items referring to speaking more than one language and the last three items focusing on goals and personal development, which may be the reason for this lack of fit. Items 2 and 3 for the introjected motivation construct are overfitting. This may be because these items are

unique in containing reference to negative feelings of shame or embarrassment. For the external motivation construct, items 1 and 2 are underfitting, whereas 4 is overfitting. Again, there is a clear split among these items, and as discussed above, some items (1, 2, and 4) refer to study and the language course itself, whereas others (3 and 5) refer to future job prospects. Finally, for the amotivation construct, items 2 and 3 are overfitting, and 4 is underfitting. These three items all refer to reasons for studying English, whereas the remaining two items concern whether students want to study English or consider it a waste of time, which may explain the lack of fit for these three items.

Overall, the results support the models of motivation proposed by SDT, with distinct constructs being measured by the BPNSFS and the LLOS. Without unidimensional measures, researchers cannot be confident in the interpretation of results (Bond et al., 2021). If questionnaire items designed to measure intrinsic motivation are actually measuring more than one construct, it is impossible to interpret how scores derived from these items may be influencing other variables. Therefore, the current study helps to establish the unidimensionality of the majority of constructs in each measure, despite the fact that the more extrinsically oriented measures of motivation from the LLOS did not perform well. The results offer encouragement for researchers interested in using SDT as a framework to understand motivation in this context.

As for the functioning of individual items, although some adaptations had already been made for this study, it seems that further refinement is needed. Although confirmatory factor analysis showed satisfactory results for both the BPNSFS and LLOS in the study by Noels et al. (2019), more detailed item-specific feedback through Rasch analysis suggests that some of the items would benefit from revision for use in the current context. The basis of motivational research is accurate measurement, and this is dependent on the quality of the individual items that comprise questionnaires. The results of this study demonstrate the individual item-level analysis that is possible and show that individual items can be refined to improve subsequent measures. Teimouri, Plonsky, and Tabandeh (2020) suggest that research can be improved by writing items for specific contexts, and the results here suggest that Rasch analysis can be instrumental in this ongoing process.

Relationships Between Measures of Motivation and TOEIC Listening and Reading

The second research question concerned the relationship between the measures of motivation and performance on the TOEIC L&R sections. The regression models using the BPNSFS had R^2 values of .06 and .07, suggesting a weak

relationship with the DVs. The models showed that autonomy was the sole predictor of performance. This supports the findings of Oga-Baldwin et al. (2017) in a Japanese elementary school context, and Tanaka (2017) in a Japanese university context. The latter study showed that autonomy was the only one of the three constructs to have a significant relationship with motivation, which in that case related to extensive reading. Tanaka argued that as extensive reading focuses on books that are simple for the reader, perceived competence was not important for her students and was not related to their performance with extensive reading. Similarly, with the participants in the current study, it may be that in order to perform well on the TOEIC L&R, autonomous individual study plays a key role. Although English classes are compulsory for most first- and second-year university students in Japan, many simply aim to pass the course in order to gain the necessary credits for graduation (Irie & Brewster, 2013). Therefore, we might hypothesize that only students who are autonomously motivated will actively seek out opportunities to study and make real gains in their test performance. Anecdotally, the authors have observed that students who are motivated to do well on the test will often buy books to study vocabulary or past test questions, displaying autonomous learning behavior. Feelings regarding classmates are unlikely to influence this kind of self study, and the questions in the BPNSFS regarding feelings of self-efficacy in English may not be related either, as they refer to speaking (item 3) and to completing difficult tasks (item 6). The reasons outlined above might explain why the constructs of perceived competence and relatedness fail to show a significant relationship with TOEIC listening scores. Kozaki and Ross (2011) found that the classroom context influenced gains in proficiency for university students in Japan, who were influenced by the attitudes of their peers, but a study by Apple, Falout, and Hill (2013) of science and engineering students in Japan showed the importance of individual motivation. This suggests that motivation may depend on specific contexts.

The two regression models for the BPNSFS and TOEIC L&R scores showed almost identical results, indicating no real difference in the influence of the motivation variables on the two different parts of the TOEIC L&R. Although the BPNSFS was adapted for the current study, the low degree of variance explained by the regression models may be the result of using a general measure that was not originally designed for language learners. In a similar way, Teimouri et al. (2020) found that items written for language learning grit were more successful in predicting behavior than a general grit measure, and the results of the regression analyses above, with only a limited amount of variance accounted for by the models, suggest that the BPNSFS,

which is a general measure used in psychology, may be improved by further revising questions to relate even more directly to language learning.

Regression models were also constructed to investigate the relationship between the LLOS and the TOEIC L&R scores. For both TOEIC L&R, introjected motivation acted as a suppressor variable, improving the overall model by accounting for the shared variance in IVs not related to the DV (Pedhazur, 1997). It is possible that in this context, feelings of obligation to study English permeate many different types of motivation at some level, and these feelings of obligation are similar to the ought-to L2 self described by Dörnyei's (2005) model of L2 motivation. There have been mixed findings in the Japanese context regarding the ought-to self. Munezane (2013) found that although students' ought-to self influenced attitudes to English, there was no direct relationship with L2 motivation. However, studies by Taguchi et al. (2009), Yashima et al. (2017), and Apple et al. (2020) all suggested a role for the ought-to self in this context. In the current study, feelings of obligation had no direct relationship with the DVs, supporting the findings of Munezane (2013), whose study was also conducted in a university context in Japan. Similarly, in the current study external motivation also had no relationship with performance on either test, which suggests that external pressures such as grades and future job prospects have limited impact on students in this university context. This is in contrast to the findings of Kozaki and Ross (2011) that the career aspirations of Japanese university students had a positive impact on growth in proficiency. Yet, for our university student participants, attempts to motivate students through rewards or external pressure do not seem to be effective. If students attend classes and do the required work, they are usually assured of passing the course, and therefore there is very little external motivation involved. It may be that because our participants were in their second year at university, they had not begun the process of looking for jobs (typically done during the third and fourth years) and therefore were perhaps not considering the potential benefits that a high level of English proficiency may bring to their future careers.

Integrated motivation does not seem to have relevance to either listening or reading, perhaps because in Japan, many people do not consider becoming a bilingual user of Japanese and English to be a realistic goal and it is therefore not expected or considered as part of their identity (Kozaki & Ross, 2011). This is also supported by research showing that some students in this context have no clear view of their ideal self and therefore cannot picture themselves as someone speaking English (Irie & Brewster, 2013). As would be expected, in the current study a complete lack of motivation to study was found to be significantly and negatively related to both listening and reading scores. On a

positive note, the Rasch analysis showed that amotivation was not high in the investigated group of students. Finally, identified motivation showed a weak but significant relationship with listening and reading, perhaps as a result of a general expectation that Japanese people should have some knowledge of English in order to play an active role in a globalized world (Apple et al., 2020).

Although the relationship of LLOS motivational variables to both reading and listening scores was generally very similar, intrinsic motivation was a significant predictor of listening but not of reading performance. We hypothesize that performing well on the listening test indicates an interest in English, as students need to actively seek out opportunities to engage with English media such as YouTube or to interact with other English speakers. This can be fun, intrinsically rewarding, and also can benefit performance in listening (Chien, Huang, & Huang, 2020). To do well on the reading test requires more traditional forms of study, such as intensive review of past papers and vocabulary, and may be less intrinsically enjoyable. We must stress that the above hypothesis is speculative, and more research is needed to explain the relationship between intrinsic motivation and listening and reading performance.

The only positive predictor of reading score from among the LLOS motivational variables was identified motivation, suggesting that students feel English is important for their personal development, and they may tend to adopt traditional means of study that involve reading. Identified motivation is related to internalized forms of motivation and therefore can be hypothesized to have a positive impact on intended effort and subsequent proficiency (Yashima et al., 2017).

Generally, although the regression models showed significant relationships between predictors and DVs, the relationships were moderate, accounting for small levels of variance with small to medium effect sizes (Plonsky & Oswald, 2014). This may be due in part to the receptive nature of the TOEIC L&R. Students who are highly motivated to study English and have reached high levels of conversational English may not score highly on this test. Vocabulary and grammar knowledge have been shown to correlate strongly with performance in English reading and listening proficiency tests (Hung, 2021; McLean, Stewart, & Batty, 2020), and vocabulary knowledge has been shown to develop through extramural exposure, particularly via engagement with online resources (Peters, Noreillie, Heylen, Bulté, & Desmet, 2019). It is possible that the students in the current study with higher autonomy orientations and greater intrinsic motivation seek out opportunities to engage with English through such resources, and this might explain why these constructs had a significant

relationship with listening and reading proficiency. Reading proficiency was positively predicted only by identified motivation, which is described as internal (Ryan & Deci, 2017). Further research is needed to investigate this relationship. It should be noted that in the motivation research literature, Dörnyei and Ryan (2015) describe a .40 correlation as large, and therefore the models here could arguably be interpreted by some as explaining a reasonable amount of variation on the TOEIC L&R.

There are several implications from the current study. For researchers, the findings of the current study demonstrate that Rasch analysis has clear benefits when evaluating measurement instruments. Rasch analysis is able to provide more information about the functioning of individual items than other methods such as factor analysis, allowing researchers to develop and revise items to improve measurement. The findings suggest that for the current context, although the BPNSFS items performed reasonably well, in order to explain more of the variance in listening and reading performance they should be revised to focus more on language learning. Items from the LLOS should also be revised to increase their effectiveness in explaining performance in this context. Future studies should attempt to establish how enhancing the identified motivation of Japanese L2 learners may lead students to extramural English study and engagement with authentic L2 input, which in turn can lead to benefits in L2 comprehension. We also hope that the current study contributes to the research literature through the inclusion of an outcome variable (Al-Hoorie et al., 2021), and that researchers can build on the findings with a wider range of DVs.

Limitations and Future Research

Although the large sample size was beneficial for statistical analyses, the cross-sectional design means that we were unable to identify any changes in motivation and instead provided a one-off snapshot of motivation. Furthermore, there was no behavioral measure directly reflecting the time that students invested in studying English, and we did not assess students' extramural engagement with the target language. Finally, the BPNSFS and LLOS were adapted for this specific context, which may have influenced their functioning in the subsequent analyses and the applicability of the findings to other contexts.

Future research should attempt to validate the measures used with a more diverse range of students, in terms of, for example, age, language proficiency, and language background. It would also be of interest to see how motivation relates to in-class behavior and engagement with tasks, and to investigate students' views on motivation and language learning. Finally, use of outcome

variables such as task engagement and speaking proficiency will allow researchers to understand how SDT relates to productive language use.

Conclusion

Motivation is an important individual difference variable influencing success in language learning (Dörnyei & Ryan, 2015), and in efforts to better understand it, SDT is increasingly used (McEown & Oga-Baldwin, 2019). Central to effective research is effective measurement, and although both the BPNSFS and LLOS are regularly used, this has been the first study to use Rasch analysis to examine their validity and reliability. Rasch analyses of both the BPNSFS and LLOS provided evidence for the unidimensionality of the individual constructs for which the questionnaires aim to elicit data. Extrinsic measures of motivation were weakest, with external motivation showing poor unidimensionality, and introjected motivation and external motivation showing poor person separation, indicating a lack of ability to measure variance in the population on these two constructs. Generally, validity and reliability problems with individual items may have been due to nuances in the wording of those items, suggesting that rewriting would be helpful. Wright maps highlighted potential issues with the common approach of summing Likert category scores to represent constructs. Regression analyses showed that motivation accounted for a small but significant amount of the variance in listening and reading test performance. External motivation had no relationship with test scores, and introjected motivation behaved as a suppressor variable, with no direct relationship to the TOEIC L&R scores, but accounting for common variance in the other predictor variables. A difference was noted between the differential impact of intrinsic motivation on listening versus reading performance, in that intrinsic motivation predicted the variance on listening scores but not on reading scores.

Greater understanding of how motivation influences student behavior could help language teachers to motivate their students. Central to this understanding is valid and reliable measurement of motivation, and in the current study we have attempted to facilitate this through the use of Rasch analysis.

Open Research Badges

This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. All data and materials that the authors have used and have the right to share are

available at https://mfr.osf.io/render?url=https%3A%2F%2Fosf.io%2Ffs86b%2Fdownload and http://www.iris-database.org. All proprietary materials have been precisely identified in the manuscript.

References

Agawa, T., & Takeuchi, O. (2016). Validating self-determination theory in the Japanese EFL context: Relationship between innate needs and motivation. *Asian EFL Journal*, *18*, 7–33.

Alexandrov, A. (2010). Characteristics of single-item measures in likert scale format. *The Electronic Journal of Business Research Methods*, *8*, 1–12.

Al-Hoorie, A. H., Hiver, P., Kim, T. Y., & De Costa, P. (2021). The identity crisis in language motivation research. *Journal of Language and Social Psychology*, *40*, 136–153. https://doi.org/10.1177/0261927/20964507

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, *38*, 665–680.

Apple, M. T. (2013). Using Rasch analysis to create and evaluate a measurement instrument for foreign language classroom speaking anxiety. *JALT Journal*, *35*, 5–28. https://doi.org/10.37546/jaltjj35.1-1

Apple, M. T., Da Silva, D., & Fellner, T. (2013). *Language learning motivation in Japan*. Bristol, UK: Multilingual Matters.

Apple, M. T., Falout, J., & Hill, G. (2013). Exploring classroom-based constructs of EFL motivation for science and engineering students in Japan. In M. T. Apple, D. Da Silva, & T. Fellner (Eds.), *Language learning motivation in Japan* (pp. 54–74). Bristol, UK: Multilingual Matters.

Apple, M. T., Hill, G., & Falout, J. (2020). The relationship between future career self images and English achievement test scores of Japanese STEM students. *IEEE Transactions on Professional Communication*, *63*, 372–385. https://doi.org/10.1109/tpc.2020.3029662

Apple, M. T., & Neff, P. (2012). Using Rasch measurement to validate the big five factor marker questionnaire for a Japanese university population. *Journal of Applied Measurement*, *13*, 276–296.

Arai, K. (2004). Naniga gaikokugo gakushusha no yaruki wo ushinawaserunoka? Douki gentai no genin to sorenitaisuru gakushusha no hannou ni taisuru shitsuteki chosa [What demotivates foreign language learners? Qualitative study of causes for demotivation and learners' reactions to demotivating experiences]. *Toyo Gakuin Daigaku Kiyo*, *12*, 39–47.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*, 6–40. https://doi.org/10.1177/0265532220927487

Benesse Educational Research and Development Institute. (2015). *Sho/Chuugakusei he no doukiduke ha hogosha no dono youna kakawarikataniyottetakamaruka* [How

do guardians impact learning motivation of elementary and junior high school students]. *Benesse Institute of Education Research Report on Learning in Elementary and Junior High School*, 1–9. Retrieved from https://berd.benesse.jp/up_images/research/Survey_on_learning_report_1.pdf

Berwick, R., & Ross, S. (1989). Motivation after matriculation: Are Japanese learners of English still alive after exam hell? *JALT Journal*, *11*, 193–210.

Bokander, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning*, *70*, 11–47. https://doi.org/10.1111/lang.12368

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). New York, NY: Routledge.

Chen, B., Vansteenkiste, M., Beyers, W., Boone, L., Deci, E. L., Van der Kaap-Deeder, J., … Verstufy, J. (2015). Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motivation and Emotion*, *39*, 216–236. https://doi.org/10.1007/s11031-014-9450-1

Chen, X., Lake, J., & Padilla, A. M. (2021). Grit and motivation for learning english among Japanese university students. *System*, *96*, 1–11. https://doi.org/10.1016/j.system.2020.102411

Chien, C., Huang, Y., & Huang, P. (2020). YouTube videos on EFL college students' listening comprehension. *English Language Teaching*, *13*, 96–103. https://doi.org/10.5539/elt.v13n6p96

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation*, *10*, 1–9. https://doi.org/10.7275/jyj1-4868

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*, 227–268.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*, 27–46. https://doi.org/10.1111/j.16000587.2012.07348.x

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Dörnyei, Z., & Csizér, K. (2002). Some dynamics of language attitudes and motivation: Results of a longitudinal nationwide survey. *Applied Linguistics*, *23*, 421–462. https://doi.org/10.1093/applin/23.4.421

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. New York, NY: Routledge.

Embertson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New York, NY: Psychology Press.

Educational Testing Service. (2008). *TOEIC test data and analysis 2007: Number of examinees and scores in FY2007*. Tokyo: Institute for International Business Corporation, TOEIC Steering Committee.

Educational Testing Service. (2013). User guide – Listening and reading. Retrieved from www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf

ETS-TOEIC (2019). The TOEIC program data & analysis 2019: 2018 Gakunendo jukenshasu to heikin sukoa [The TOEIC program data & analysis: Academic year 2018 test takers and average scores]. Retrieved from https://www.iibc-global.org/library/default/iibc/press/2019/p118/pdf/2019DAA.pdf

Falout, J., Elwood, J., & Hood, M. (2009). Demotivation: Affective states and learning outcomes. *System*, *37*, 403–417. https://doi.org/10.1016/j.system.2009.03.004

Field, A. P. (2009). *Discovering statistics using SPSS*. London, UK: Sage.

Hiromori, T. (2006). *Gaikokugo gakushusya no doukizuke wo takameru riron to jissen* [Theory and practice to improve foreign language learners]. Tokyo: Taga-Shuppan.

Hiver, P., & Al-Hoorie, A. H. (2020). Reexamining the role of vision in second language motivation: A preregistered conceptual replication of You, Dörnyei, and Csizér (2016). *Language Learning*, *70*, 48–102. https://doi.org/10.1111/lang.12371

Hung, T. H. (2021). Exploring the relationships between various dimensions of receptive vocabulary knowledge and L2 listening and reading comprehension. *Language Testing in Asia*, *11*(20), https://doi.org/10.1186/s40468-021-00131-8

Irie, K., & Brewster, D. R. (2013). One curriculum, three stories: Ideal L2 self and L2-self-discrepancy profiles. In M. T. Apple, D. Da Silva, & T. Fellner (Eds.), *Language learning motivation in Japan* (pp. 110–128). Bristol, UK: Multilingual Matters.

Kikuchi, K. (2009). Listening to our learners' voices: What demotivates Japanese high school students? *Language Teaching Research*, *13*, 453–471.

Kikuchi, K., & Sakai, H. (2009). Japanese learners' demotivation to study English: A survey study. *JALT Journal*, *31*(9), 183–204.

Knoch, U., & McNamara, T. (2015). Rasch analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 275–304). New York, NY: Routledge.

Kormos, J., & Csizér, K. (2008). Age-related differences in the motivation of learning English as a foreign language: Attitudes, selves, and motivated learning behavior. *Language Learning*, *58*(2), 327–355. https://doi.org/10.1111/j.1467-9922.2008.00443.x

Kormos, J., Kiddle, T., & Csizér, K. (2011). Systems of goals, attitudes, and self-related beliefs in second language learning motivation. *Applied Linguistics*, *32*, 495–516. https://doi.org/10.1093/applin/amr019

Kozaki, Y., & Ross, S. J. (2011). Contextual dynamics in foreign language learning motivation. *Language Learning*, *61*, 1238–1354. https://doi.org/10.1111/j.1467-9922.2011.00638.x

Leeming, P. (2017). A longitudinal investigation into English speaking self-efficacy in a Japanese language classroom. *Asian-Pacific Journal of Second and Foreign Language Education*, *2*, https://doi.org/10.1186/s40862-017-0035-x

Leeming, P., & Harris, J. (2021a). *Surveys. Materials from "Measuring foreign language students' self-determination: A Rasch validation study"* [Questionnaire]. IRIS Database, University of York, UK. https://doi.org/10.48316/668t-rb20

Leeming, P., & Harris, J. (2021b). *Survey data. Datasets from "Measuring foreign language students' self-determination: A Rasch validation study"* [Dataset]. IRIS Database, University of York, UK. https://doi.org/10.48316/v4va-5g80

Linacre, J. M. (2006). Data variance explained by Rasch measures. *Rasch Measurement Transactions*, *20*, 1045–1054.

Linacre, J. M. (2020a). *WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis* [Computer software]. Available from Retrieved from https://www.winsteps.com/winsteps.htm

Linacre, J. M. (2020b). *A user's guide to WINSTEPS: Rasch-model computer program (4.7.0)*. Chicago, IL: MESA Press.

Lou, N. M., Chaffee, K. E., Vargas Lascano, D. I., Dincer, A., & Noels, K. A. (2018). Complementary perspectives on autonomy in self-determination theory and language learner autonomy. *TESOL Quarterly*, *52*, 210–220. https://doi.org/10.1002/tesq.403

McEown, M. S., & Oga-Baldwin, W. L. Q. (2019). Self-determination for all language learners: New applications for formal language education. *System*, *86*, 102124. https://doi.org/10.1016/j.system.2019.102124

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*, 389–411. Retrieved from https://doi.org/10.1177/0265532219898380

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*, 555–576. https://doi.org/10.1177/0265532211430367

MEXT. (2011). Five proposals and specific measures for developing proficiency in English for international communication. Retrieved from https://www.mext.go.jp/component/english/__icsFiles/afieldfile/2012/07/09/1319707_1.pdf

Munezane, Y. (2013). Motivation, ideal L2 self and valuing of global English. In M. T. Apple, D. Da Silva, & T. Fellner (Eds.), *Language learning motivation in Japan* (pp. 152–168). Bristol, UK: Multilingual Matters.

Nicholson, S. J. (2015). Evaluating the TOEIC® in South Korea: Practicality, reliability and validity. *International Journal of Education*, *7*, 221–233. https://doi.org/10.5296/ije.v7i1.7148

Nishimura, T., & Suzuki, T. (2016). Basic psychological need satisfaction and frustration in Japan: Controlling for big five personality traits. *Japanese Psychological Research*, *58*, 320–331. https://doi.org/10.1111/jpr.12131

Noels, K. A., Lou, N. M., Vargas Lascano, D. I., Chaffee, K. E., Dincer, A., Zhang, Y. S. D., & Zhang, X. (2019). Self-determination and motivated engagement in language learning. In M. Lamb, K. Csizér, A. Henry, & S. Ryan (Eds.), *The Palgrave handbook of motivation for language learning* (pp. 95–115). Cham, Switzerland: Macmillan. https://doi.org/10.1007/978-3-030-28380-3

Noels, K. A., Pelletier, L. G., Clément, R., & Vallerand, R. J. (2000). Why are you learning a second language? Motivational orientations and self-determination theory. *Language Learning*, *50*, 57–85. https://doi.org/10.1111/0023-8333.00111

Noels, K. A., Vargas Lascano, D. I., & Saumure, K. (2019). The development of self-determination across the language course: Trajectories of motivational change and the dynamic interplay of psychological needs, orientations, and engagement. *Studies in Second Language Acquisition*, *41*, 821–851. https://doi.org/10.1017/S0272263118000189

Oga-Baldwin, W. Q., Nakata, Y., Parker, P., & Ryan, R. M. (2017). Motivating young language learners: A longitudinal model of self-determined motivation in elementary school foreign language classes. *Contemporary Educational Psychology*, *49*, 140–150. https://doi.org/10.1016/j.cedpsych.2017.01.010

Pedhazur, E. (1997). *Multiple regression in behavioral research: Explanation and prediction*. Fort Worth, TX: Wadsworth.

Peters, E., Noreillie, A.-S., Heylen, K., Bulté, B., & Desmet, P. (2019). The impact of instruction and out-of-school exposure to foreign language input on learners' vocabulary knowledge in two languages. *Language Learning*, *69*, 747–782. https://doi.org/10.1111/lang.12351

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. https://doi.org/10.1111/lang.12079

Powers, D. E., & Powers, A. (2015). The incremental contribution of TOEIC listening, reading, speaking, and writing tests to predicting performance on real-life English language tasks. *Language Testing*, *32*, 151–167. https://doi.org/10.1177/0265532214551855

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, *65*(S1), 37–75. https://doi.org/10.1111/lang.12112

Raîche, G. (2005). Critical eigenvalue sizes (variances) in standardized residual principal components analysis. *Rasch Measurement Transactions*, *19*, 1012. https://doi.org/10.1177/1534484310380331

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Denmark Paedagogiske Institut.

Reio, T. G. (2010). The threat of common method variance bias to theory building. *Human Resource Development Review*, *9*, 405–411.

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. New York, NY: Guildford Press.

Salzberger, T. (2010). Does the Rasch model convert an ordinal scale into an interval scale? *Rasch Measurement Transactions*, *24*, 1273–1275.

Sharma, R., Yetton, P., & Crawford, J. (2009). Estimating the effect of common method variance: The method-method pair technique with an illustration from TAM research. *MIS Quarterly*, *33*, 473–490.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, *1*, 199–218.

Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 316–327). Greenwich, CT: Ablex.

Someya, F. (2017). A survey on the relationship between motivation and listening comprehension ability of 1st-grade junior high school students: Focusing on Self-Determination Theory. *Journal of the Chubu English Lanugage Education Society*, *46*, 179–183.

Someya, F. (2018). A survey on the motivational differences between elementary school pupils and junior high school students: Focussing on Organismic Integration Theory. *The Faculty Journal of Komazawa Women's University*, *25*, 259–271.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, MA: Pearson/Allyn & Bacon.

Taguchi, T., Magid, M., & Papi, M. (2009). The L2 Motivational Self System among Japanese, Chinese, and Iranian learners of English: A comparative study. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 66–97). Bristol, UK: Multilingual Matters.

Tanaka, M. (2017). Factors affecting motivation for short in-class extensive reading. *The Journal of AsiaTEFL*, *14*, 98–113. https://doi.org/10.18823/asiatefl.2017.14.1.7.98

Teimouri, Y., Plonsky, L., & Tabandeh, F. (2020). L2 grit: Passion and perseverance for second language learning. *Language Teaching Research*, 1–48. https://doi.org/10.1177/1362168820921895

Ushioda, E. (2013). Foreign language motivation research in Japan: An 'insider' perspective from outside Japan. In M. T. Apple, D. Da Silva, & T. Fellner (Eds.), *Language learning motivation in Japan* (pp. 1–14). Bristol, UK: Multilingual Matters.

Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231–251. https://doi.org/10.1037/1082-989X.3.2.231

Wang, C., Kim, D., Bai, R., & Hu, J. (2014). Psychometric properties of a self-efficacy scale for English language learners in China. *System*, *44*, 24–33. https://doi.org/10.1016/j.system.2014.01.015

Waugh, R. F., & Chapman, E. S. (2005). An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: What is the difference? Which method is better? *Journal of Applied Measurement*, *6*, 80–99.

Weaver, C. (2005). Using the Rasch model to develop a measure of second language learners' willingness to communicate within a language classroom. *Journal of Applied Measurement*, *6*, 396–415.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wolfe, E. W., & Smith, E. V. (2007a). JrInstrument development tools and activities for measure validation using Rasch models: Part I – Instrument development tools. *Journal of Applied Measurement*, *8*, 97–123.

Wolfe, E. W., & Smith, E. V. (2007b). JrInstrument development tools and activities for measure validation using Rasch models: Part II – Validation activities. *Journal of Applied Measurement*, *8*, 204–234.

Wright, B. D. (1996a). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, *3*, 3–24. https://doi.org/10.1080/10705519609540026

Wright, B. D. (1996b). Reliability and separation. *Rasch Measurement Transactions*, *9*, 472.

Yamauchi, H., & Tanaka, K. (1998). Relation of autonomy, self-referenced beliefs, and self-regulated learning among Japanese children. *Psychological Reports*, *82*, 803–816. https://doi.org/10.2466/pr0.1998.82.3.803

Yashima, T., Nishida, R., & Mizumoto, A. (2017). Influence of learner beliefs and gender on the motivating power of L2 selves. *The Modern Language Journal*, *101*, 691–711. https://doi.org/10.1111/modl.12430

Yashima, T., Zenuk-Nishide, L., & Shimizu, K. (2004). The influence of attitudes and affect on willingness to communicate and second language communication. *Language Learning*, *54*, 119–152. https://doi.org/10.1111/j.1467-9922.2004.00250.x

Yun, S., Hiver, P., & Al-Hoorie, A. H. (2018). Academic buoyancy. *Studies in Second Language Acquisition*, *40*, 805–830. https://doi.org/10.1017/s0272263118000037

Zhang, S. (2006). Investigating the relative effects of persons, items, sections, and languages on TOEIC score dependability. *Language Testing*, *23*, 351–369. https://doi.org/10.1191/0265532206lt332oa

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Basic Psychological Needs Satisfaction and Frustration Scale (BPNSFS).
Appendix S2. Language Learning Orientation Scale (LLOS).
Appendix S3. Category Structure Functioning for BPNSFS.
Appendix S4. Category Structure Functioning for LLOS.
Appendix S5. Descriptive Statistics for TOEIC Listening and Reading.

Appendix S6. Descriptive Statistics for Motivational Variables.
Appendix S7. Correlations for the TOEIC L&R and Motivation Constructs.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Self-determination theory measures of motivation: Are they valid and do they predict English reading and listening performance?

What This Research Was About and Why It Is Important

Research on motivation often uses questionnaires, but we need to be confident that these questionnaires are valid (i.e., measure what they claim to measure). Two questionnaires commonly used in self-determination theory (SDT) research are the Basic Psychological Needs Satisfaction and Frustration Scale (BPNSFS) and the Language Learning Orientations Scale (LLOS). In this study, data from these questionnaires, administered to learners of English in a Japanese university, were analyzed using the Rasch model for measurement to establish their validity. Also, a regression analysis examined whether motivation, as measured by these questionnaires, predicted English listening and reading proficiency. It was found that the questionnaires satisfied many of the requirements of the Rasch model, thus demonstrating their validity, but showed some items needed revision. Results also showed that autonomy (a sense of freedom regarding choices about learning) had a significant relationship with English listening and reading proficiency; intrinsic motivation (a sense of fulfilment and enjoyment from learning) was related to English listening but not reading; and competence (self-perceived ability) and external motivations (e.g, studying for material rewards such as grades or future employment) seemed not to have any relationship with listening or reading scores.

What the Researchers Did

- The Basic Psychological Needs Satisfaction and Frustration Scale (BPNSFS) and the Language Learning Orientations Scale (LLOS) were given to 600 second-year non-English major university students in Japan studying English as a foreign language.
- The same students completed the TOEIC Listening and Reading proficiency tests.
- Each of the questionnaires was analyzed using the Rasch model for measurement.
- Results from both motivation questionnaires were used to predict scores on the English proficiency test.

What the Researchers Found

- Rasch analysis showed that the questionnaires measured the hypothesized constructs appropriately.
- There were some problems with measurement of external forms of motivation.
- Although the constructs were largely valid, most also had issues that seemed to be linked to wording of individual questionnaire items.
- Autonomy was a significant predictor of listening and reading proficiency.
- Intrinsic motivation predicted listening proficiency.
- External forms of motivation had no relationship with reading or listening proficiency.

Things to Consider

- Rasch analysis offered detailed information about whether each construct measured by the questionnaires was indeed *one* construct (versus multiple constructs), and also about how individual items contribute to the data for each construct.
- Questionnaires can be improved by rewriting items based on the results of Rasch analysis.
- To improve English listening and reading proficiency, external rewards or threats may not be as useful as enhancing students' internal forms of motivation (e.g., satisfaction from learning).
- Intrinsic motivation, which was associated with listening proficiency, may have been caused by increased engagement with English outside of class.

Materials and data: Materials and data are publicly available at IRIS (https://www.iris-database.org) and OSF (https://mfr.osf.io/render?url = https://3A/2F/2Fosf.io/2Ffs86b/2Fdownload).

How to cite this summary: Leeming, P., & Harris, J. (2022). Self-determination theory measures of motivation: Are they valid and do they predict English reading and listening performance? *OASIS Summary* of Leeming & Harris (2022) in *Language Learning*. https://oasis-database.org