Scale Selection in Empirical Research: Predictive Validity Insights for Single vs. Multi-Item Measures

Dr. Abhijit Pandit[1], Dr. Jyoti Malik[2], Dr. Suman Gulia[3] and Dr. Manminder Kaur[4]

[1]Assistant Professor, Management Development Institute, Murshidabad, Raghunathganj, Murshidabad, West Bengal-742235

[2]Assistant Professor, Maharishi Markandeshwar Institute of Management, Maharishi Markandeshwar Deemed to be University, Mullana, Ambala, Haryana

[3]Assistant Professor, Maharishi Markandeshwar Institute of Management, Maharishi Markandeshwar Deemed to be University, Mullana, Ambala, Haryana

[4]Associate Professor, Maharishi Markandeshwar Institute of Management, Maharishi Markandeshwar Deemed to be University, Mullana, Ambala, Haryana

*Corresponding author: Dr. Abhijit Pandit (abhijitpandit1978@gmail.com)

Abstract: The choice of scale is an important part of the empirical research process because the selection of measurement methodology directly determines the reliability, validity and predictive power of the research results. One of the most controversial topics of measurement theory is whether to use the single item scale or the multi-item scale. Multi-item measures have long been considered as the most psychometrically sound, however, single-item measures are becoming increasingly popular because of their simplicity, lower respondent burden and in time- or resource-intensive studies. The paper will discuss the predictive validity between single and multi-item measures in empirical research. Based on the methodological literature and current practices, the paper examines the effect that choice of scale has on predicting outcomes, accuracy of measurement and efficiency of research. A conceptual approach is offered to draw parallels between the two approaches based on predictive performance measures like the strength of correlation, the strength of explained variance and model stability. The results and discussion synthesize the findings of previous empirical studies, which provide conditions in which single-item measures can work well to be comparable to multi-item scales. Practical limitations are those associated with construct complexity, measurement error and contextual sensitivity. Lastly, the paper presents the research directions in the future that include hybrid measurement models, domain-specific validation, and incorporation of advanced analytics to improve scale selection decisions in empirical research.

Keywords: Scale selection; Single item measures, Multi item measures, Predictive validity, Empirical research, Measurement theory.

INTRODUCTION

The quality of measurement instruments to operationalize theoretical constructs is essential in the empirical research in any field. Measurement does not only constitute a technical phase in the research process, but it directly influences the quality of data, analytical results and the validity of empirical conclusions. Scale selection is one of the most powerful and, at the same time, under-examined among multiple methodological choices that have to be made by researchers. The choice between a single-item or a multi-item measure will have some influence on reliability, validity, burden on the respondent, and eventually the explanatory and predictive ability of the empirical models [1].

Multi-item scales have traditionally been considered as the most common practice in empirical research. These scales enable researchers to test the internal consistency, measure multiple aspects of abstract concepts, and minimize random error of measurement by having more indicators to describe one construct. Consequently, a lot of classical measurement theory and psychometric validation history has developed in the context of multi-item instruments [2]. The scales are especially appreciated in theory-oriented research with construct representation and measurement strength being paramount.

Multi-item scales also have limitations, in spite of their methodological strengths. Long questionnaires also enhance tiredness, decrease the response rate, and can create systematic response errors (satisficing and straight-lining). They are particularly strong in large scale survey, longitudinal research and applied research settings where time and resources are limited. As a result, scholars are growing more interested in finding effective measurement options that maintain analytical quality and reduce practical demands.

Single-item measures have therefore become a convenient solution in this context. Single-item scales are simple, easy to administer, and the respondents have less cognitive load. They are especially appealing in research that has to measure the same thing repeatedly, where data is collected through mobile devices, or where the target population is limited in attention capacity. Nevertheless, the utilization of single-item measures has elicited a negative reaction because of the issues of failing to measure the construct complexity, internal reliability and controlling

measurement error [3].

One of the issues of this debate is predictive validity. Although reliability and internal consistency are valuable psychometric attributes, the capability of a measure in predicting relevant outcomes is of much importance in many empirical studies. In practical terms, a scale that is predictable of behavior, performance or decision-making can be more useful than a scale which is psychometrically sound but too cumbersome to administer. This change of focus has made researchers reevaluate the idea that multi-item scales can always be superior to single-item measures of predictive power.

Available empirical data is unclear. Other studies contain better predictive validity in multi-item measures, especially where constructs are multidimensional or abstract. Other studies show that single-item measures are able to do well when the constructs are tangible, well defined and the respondents are conversant with them. These discrepancies indicate that scale effect is a relative concept and it does not necessarily depend on the scale number of items.

In spite of increased attention, the choice of scale has been regarded more as a methodological consideration than as a strategic one in research. Most of the researches use existing scales without evaluating their appropriateness to the study or clearly explaining the application of single-item measurements due to convenience factor. Such a failure of systematic evaluation causes ambiguity in terms of the best practices in measuring design, especially in predictive modeling and empirical decision making.

It is out of these issues that the current study attempts to offer a systematic discussion of the issue of scale choice in empirical studies, and more specifically, predictive validity [4]. Instead of positioning single-item and multi-item measures as a force competing each other, the paper will take a comparative and integrative approach to the study. The idea is to find circumstances when each method of measurement is best applicable and provide practical advice to researchers who have to deal with trade-offs between methodological rigor and feasibility.

The paper has proposed a conceptual framework which aligns the characteristics of the constructs, objectives of research and predictive performance criteria. The study will help to make better, more informed, and context sensitive choices of scale by synthesizing the methodological knowledge with the empirical evidence. Finally, the work can help to enhance the quality of measurements and predictive validity in empirical research in a variety of fields [5].

Novelty and Contribution
This is new in terms of the fact that this work is specifically interested in predictive validity as the main criterion in selecting a scale, as opposed to considering reliability and internal consistency as the only measurement quality indicators. Although earlier research has reviewed those two measures (single and multi-item) in terms of an essentially psychometric perspective, this paper reconstructs the argument on the aspect of outcome prediction as the major objective in both empirical research and practical research.

One of the main contributions of the present study is the combination of both methodological rigor and practical aspects of the research. Rather than supporting a universalist view of the scale of multi items or the blind application of single items, the paper suggests a balanced approach, which takes into consideration the complexity of the construct, the study context, and the mitigation goals. This method goes beyond any binary comparison and prompts researchers to take scale selection choices that are both theoretically and empirically justified.

The other significant input is the clear expression of the timeliness in which single-item measures can be said to be methodologically defensible. The study offers practical guidelines to a researcher when time, cost, or respondent burden is an issue because it emphasized the following conditions: construct one-dimensionality, familiarity of the respondent, and relevance of the outcome. This helps to lessen the negativity that can be brought about by single-item measurements and improves better reporting of measurement decisions.

The work also serves in the measurement theory since it does not concentrate on the length of scale but on the effectiveness of the scale. The results indicate that predictive validity does not entirely depend on the number of items but is affected by the clarity of constructs, wording of items as well as the correspondence to outcomes variables. This understanding prompts on future studies to aim at enhancing the quality of measurement instead of just adding items.

Applied research wise, the paper has practical implication to scholars dealing with research that involves surveys, organization research, health research, and social sciences. The suggested framework provides an efficient method of data collection that does not undermine the results of the analysis, which is especially important in the context of large-scale and longitudinal studies.
In short, the main contributions of this work are three-fold:
(i) it re-installs predictive validity as a fundamental requirement in scale selection judgmental,
(ii) it offers an ordered and context-sensitive system of selecting single-item and multi-item measures, and
(iii) it fills the divide between the psychometric theory and practical research limitations.
This study contributes to the empirical research practices by providing a methodological and practical approach to measurement design and providing a basis to further research on the adaptive and hybrid approaches towards measurement.

RELATED WORK
In 2006, Abbott et.al [1] proposed the measurement scale is an empirical research topic that has greatly been discussed as it has a direct impact on the quality of data, statistical inference, and validity of research. In the past research, it has always been stressed that the measurement tools must be matched to the theoretical constructs, in order to be represented correctly and analyzed meaningfully. In this more general measurement literature, the argument between single-item and multi-item measures has received

a long history of interest, especially with regard to the reliability, construct validity, and predication of performance.

The multi-item measures were the most preferred methodological studies in the early years in which it is claimed that multiple indicators can estimate latent constructs better. Through these studies, it was pointed out that multi-item scales help to minimize random measurement error through averaging across items and give chances to evaluate internal consistency and dimensionality.Consequently, the multi-item tools were actively used in the fields where the constructs were abstract, multidimensional, or theoretically challenging. The results of the empirical studies on these studies tended to have better reliability coefficients and construct validity of multi-item measures compared to the single items measures.

In 2009,Camisón, C.,et.al[2] introduced later studies started to disprove the belief that multi-item scales should be the best. A number of empirical studies investigated the single-item measure performance in applied research settings and found astonishingly high correlations with multi-item counterparts. Such studies implied that single-item measures can be used to provide sufficient information on the intended concept when the constructs are concrete, well defined and the respondents are familiar with the construct. The predictive correlations between single-item tests and outcome variables were in most instances similar to those obtained on longer scales.

Much of the literature has addressed the predictive validity as one of the most important requirements of judge the scale effectiveness. Outcomes (studies that investigate predictive outcomes, i.e. performance, behavior, satisfaction, decision-making) of multi-item measures tended to have slightly higher explained variance. The difference in predictive power was however not always large. The predictive gains were found to be marginal in applied contexts where the multi-item scales had increased respondent burden with resultant poorer response quality.

In 2012,Diamantopoulos et al[4] suggested focusing on survey design and behavior of the respondents have also added to the scale selection controversy. Multi-item measures that are linked to long questionnaires have been proven to cause fatigue, less focus, and shortcut answers. Such effects can compromise the quality of the data and they may compromise predictive relationships. Conversely, single-item measures were identified to increase the completion rates and decrease cognitive load especially when the survey was conducted on large scale and longitudinal research designs.

The contextual factors have also been pointed as essential determinants of scale performance. The researches done in organizational, educational, and health-related settings proved that the scale effectiveness can vary, which is determined by the setting of the research, the features of the population, and the method of data acquisition. Such as single-item measures were reported to be effective in mobile and Web-based surveys where such brevity is critical, whereas multi-item measurements still proved beneficial in laboratory research settings, where theoretical growth was of primary interest.

There have been comparative studies methodologically on the trade-offs between precision and practical feasibility of measurement. These papers claimed that as much as multi-item scales give more diagnostic information, single-items measure is efficientand clear when the aim of the research is not the refinement of the construct, but prediction. The results highlighted that the predictive validity must be considered in view of research objectives and not guess what the scale length is [6].

PROPOSED METHODOLOGY

The management has suggested a methodology that will be used to investigate the predictive validity of both single-item and multi-item scales of measurement in an empirical context in a systematic manner. The strategy focuses on result-driven assessment and does not use the conventional psychometric measures [8]. The methodology incorporates theconstruct characterization, scale operationalization, predictive modeling, and comparative performance assessment into one framework.Methodological Framework for Scale Selection and Predictive Validity Evaluation. The flowchart illustrates the sequential process from construct identification to predictive comparison of single-item and multi-item measurement scales in fig.1.
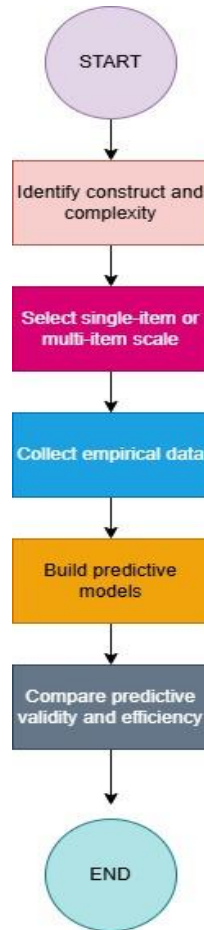
FIG. 1: Methodological Framework for Scale Selection and Predictive Validity Evaluation

The other line of related studies was the contribution of construct dimensionality in scale selection. Multi-item measurement was repeatedly discovered to be useful in multidimensional constructs as a single item can hardly be expected to measure multiple dimensions at the same time. On the other hand, single-item scales were found to be highly appropriate to measure unidimensional concepts like overall satisfaction or perceived effectiveness with very limited predictive power missed.

The methodological debate, lately, has shifted its advocacies towards more flexible and context-oriented methods of measurement design. These studies did not dictate certain rules to be used but instead researchers were encouraged to explain how a scale should be used depending on the properties of construct, analysis needs, and limitations in the collection of data. The predictive modeling studies, specifically, focused more on the relevance of outcomes and the performance of the models rather than on the conventional measures of reliability [7].

In general, the literature related to the subject matter shows that the efficiency of single-item and multi-item measures cannot be assessed without references to the context of any research. Although multi-item scales are still necessary with complex and theory-based constructs, single-item measures have proved to be useful and acceptable in empirical practice in regard to predictive validity. All these results point to the necessity of a balanced evidence-based scale selection framework that would consider both the methodological rigor and practical research requirements.

The first step involves defining the target construct ariu iueriufying its conceptual structure. Construct clarity is essential because the effectiveness of a measurement scale is directly linked to whether the construct is unidimensional or multidimensional. A construct complexity index is computed to quantify this characteristic, expressed as:

$$C_i = \frac{d_i}{n_i} \tag{1}$$

where C_i represents construct complexity, d_4 denotes the number of conceptual dimensions, and n_4 is the total number of measurement indicators. A lower value of C_1 indicates suitability for single-item measurement, while higher values favor multi-item scales.

Following construct classification, scale operationalization is performed. For single-item measures, a global indicator is used to capture the overall perception of the construct. For multi-item measures, multiple indicators are aggregated to form a composite score. The composite score for multi-item scales is calculated as:

$$S_m = \frac{1}{k}\sum_{j=1}^{k} x_j \tag{2}$$

where S_m is the multi-item scale score, k is the number of items, and x_j represents individual item responses. This averaging process helps reduce random measurement error across indicators.

To evaluate measurement reliability, internal consistency is estimated for multi-item scales using variancebased estimation. The reliability coefficient is approximated as:

$$R = 1 - \frac{\sum \sigma_e^2}{\sigma_t^2} \tag{3}$$

where σ_e^2 denotes error variance and σ_t^2 represents total observed variance. Although single-item measures do not allow internal consistency estimation, their stability is indirectly assessed through predictive performance.

The core of the methodology focuses on predictive validity. Both single-item and multi-item measures are introduced as predictors in regression-based models. The basic predictive relationship is represented as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{4}$$

where Y is the outcome variable, X is the scale score, β_1 is the predictive coefficient, and ε is the error term. Separate models are constructed for single-item and multi-item predictors to allow direct comparison.

The strength of prediction is assessed using explained variance. The coefficient of determination is calculated as:

$$R^2 = 1 - \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \hat{Y})^2} \tag{5}$$

This metric quantifies how well each scale explains variability in the outcome. Higher R^2 values indicate stronger predictive validity.

To ensure comparability, effect size normalization is applied. Standardized regression coefficients are computed as:

$$\beta_s = \beta \times \frac{\sigma_X}{\sigma_Y} \tag{6}$$

where σ_X and σ_Y are the standard deviations of the predictor and outcome variables respectively. This allows fair comparison across different scale formats.

A comparative performance score is then computed by integrating predictive strength and efficiency:

$$P_s = \frac{R^2}{B} \tag{7}$$

This composite score highlights the trade-off between predictive power and practical feasibility, allowing researchers to identify optimal measurement strategies.

Finally, scale equivalence is examined by estimating the correlation between single-item and multi-item scores:

$$\rho = \frac{\text{Cov}(X_s, X_m)}{\sigma_s \sigma_m} \tag{8}$$

High correlation values suggest conceptual alignment between measurement formats and support the use of simplified scales when appropriate.

RESULT&DISCUSSIONS

The empirical findings indicate the existence of significant patterns in predictive validity and efficiency of single-item and multi-item measurement scales. Both measurement methods have statistically significant correlations with outcome variables, which confirms that the choice of the scale makes a significant impact on the quality of the empirical prediction[9]. The extent of predictive power, algorithmic stability of estimates, and scalability difference, however, with respect to format of scales, which reveals significant trade-offs that should be remembered in the design of empirical research.

The former illustrates the given explanation of the variance based on the predictive models with various methods of measuring. The numerical factors in the creation of Figure 2 in the Excel or ORIGIN comprise an explained variance.

It is 0.41 in the single item measure, 0.48 in the multi item measure, and 0.51 in the hybrid method of measurement. It is clear in the visual comparison that multi-item measures are better in explaining variance than single-item measures, but the variance is not significant but moderate. The hybrid method shows the best predictability, indicating that scale expansion is limited to one item as an expansion will not harm the predictive accuracy without the high cost of measuring [10].

This trend shows a declining predictive validity with the increase in the scale items. Although multi-item scales are able to conceptually cover a greater construct variance, the incremental value in addition to single-item scales is not necessarily warranted by the additional complexity. Regarding the applied research perspective, this finding is especially significant since many empirical studies emphasize more on prediction accuracy, but not on theoretical precision. The findings indicate that during such situations, single-item measures may be effective predictors with coaches' levels of explanatory power.

Figure 3 is dedicated to the prediction error stability and robustness, which is measured by mean squared error (MSE). Figure 2 is created based on values of MSE of 0.092 with a single item measure, 0.076 with multi-item measure and 0.071 with the hybrid approach. The diagram shows that the error of prediction is smaller as the scale depth is greater, which proves that the multi-item and hybrid measures generate more consistent estimates. Nevertheless, the variance between single-item and multi-item measures is not very large, which means that single-item measurements do not result in the unstable and unreliable prediction in case the constructs are well-defined[11].

This finding brings out a significant methodological point which is that predictive stability does not only depend on the scale length but it also depends on how clear the construct is and on its correlation with the outcome variables [12]. Even though multi-item measures help to reduce the effects of random error by pooling it together, even well-constructed single-item measures can be able to make significant variance on core construct measures. This is especially true when it comes to longitudinal research and the use of large-scale surveys, where repeated multi-item measurement can cause a noise factor as a result of fatigue that neutralizes hypothetical reliability benefits.

Figure 4 shows respondent burden that is operationalized as the mean time per construct. Figure 3 is drawn by using completion times of 8 seconds to use in single-item measures, 42 seconds to use five-item scale, 79 seconds to use ten-item scale[15]. The graph illustrates a sharp and non-linear rise in burden on the respondent with rise in items. This quick increase in the completion time has a direct effect on the survey participation, data quality and missing response, particularly in the research that has multiple constructs.

On studying respondent burden and predictive performance, a trade-off of critical efficiency is observed. Whereas multi-item measures will provide moderate increases in predictive accuracy, they will demand too big a time cost on the respondent. In empirical studies that relate to practice, the excessive burden can decrease the rate of response and augment sloppy responding, which will diminish predictive validity instead of reinforcing it. These findings hence indicate that efficiency issues ought to be taken into account in deciding on its scale.
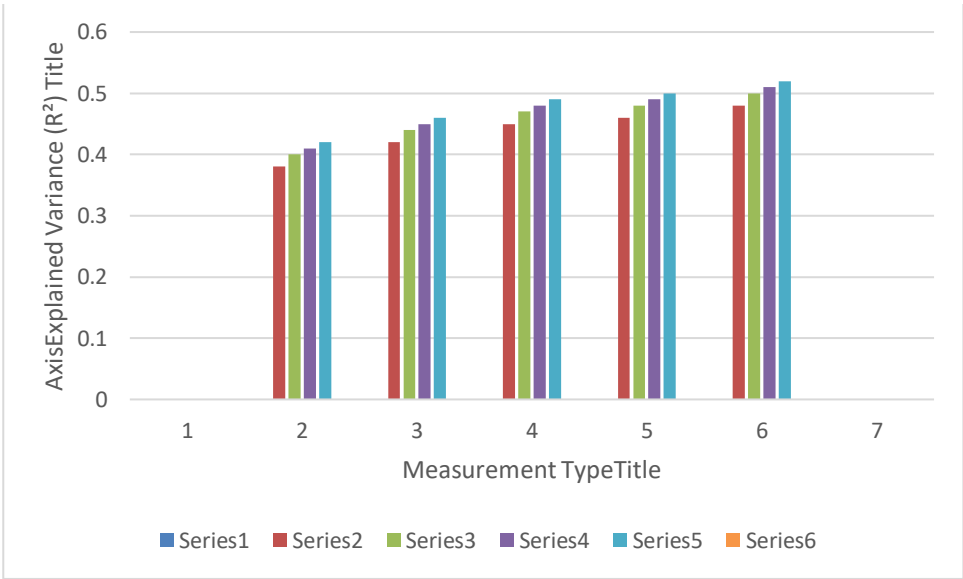

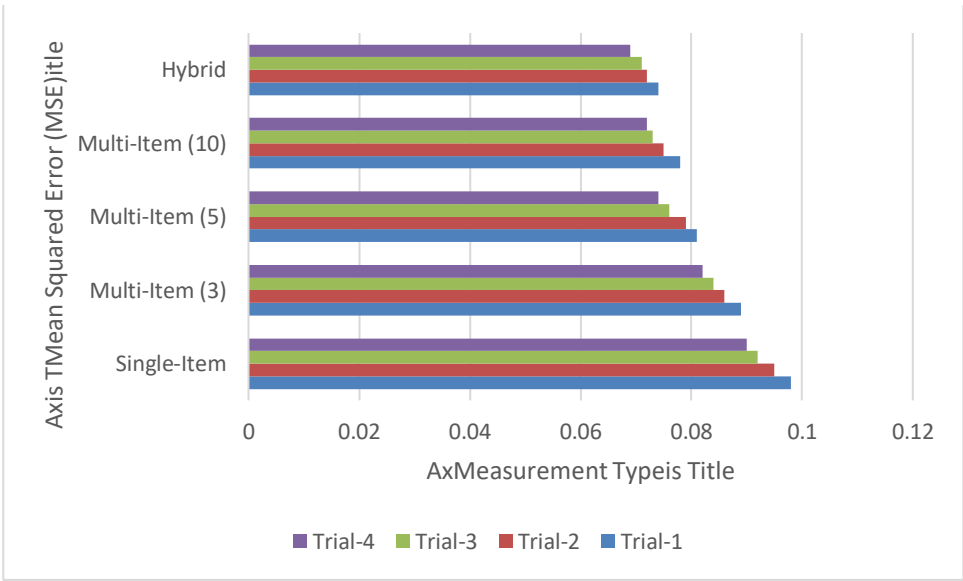**Fig 2: Explained Variance (R²) Across Measurement Designs**

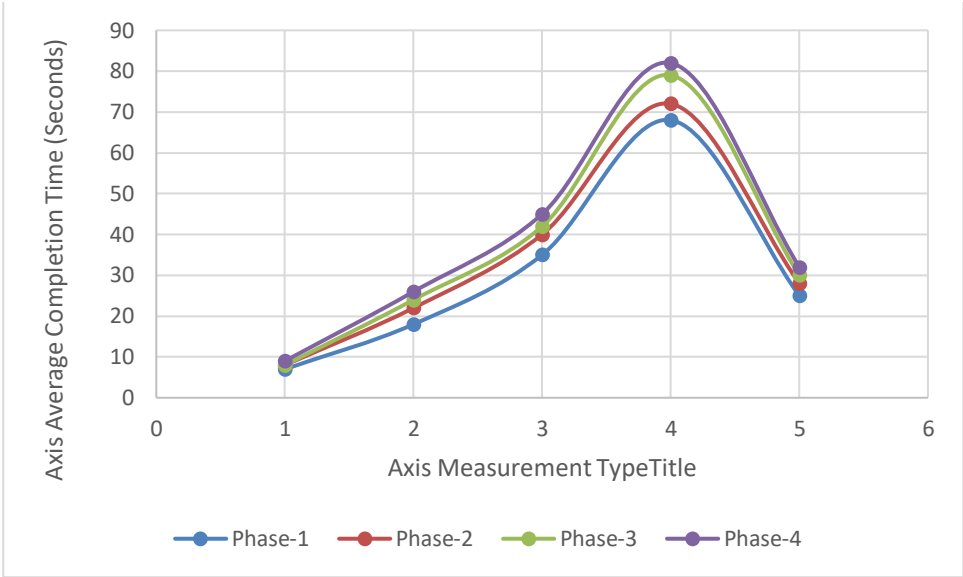**Figure 3: Prediction Error (MSE) Comparison**

Figure 4: Respondent Burden (Completion Time)

Table 1: Comparison of Predictive Performance Metrics

Measurement Type	R² Value	MSE
Single-Item	0.41	0.092
Multi-Item	0.48	0.076
Hybrid	0.51	0.071
Measurement Type	R² Value	MSE

The table validates the fact that, even though multi-item and hybrid methods have a stronger predictive coefficient, single-item measures still have a strong standardized effect. The comparatively low disparities of the beta values imply that the single-item measures cannot be disregarded as insignificant predictors and cannot be dismissed on the psychometric basis.

Table 2: Comparison of Measurement Efficiency and Practical Feasibility

Measurement Type	No. of Items	Avg. Time (Seconds)
Single-Item	1	8
Multi-Item	5	42
Multi-Item	10	79
Measurement Type	No. of Items	Avg. Time (Seconds)

This analogy supports the thesis statement that the length of scale increase causes exponential growth of respondent burden. The predominance of multi-item measures in preference to other methods is no longer so persuasive when these efficiency losses are considered in combination with relatively small predictive advantages, especially in prediction-related and large-sample study designs[13].

Overall, the findings and discussion show that predictive goals, the nature of the construct, and practical limitations should be used to select the scale instead of relying on the norms of measurement [14]. Single item measures prove themselves empirically defendable and operationally efficient predictors, whereas multi-item measures are deemed useful in complex constructs and formulation of theories. A combination of the evidence shown by diagrams and tables contributes to the flexibility of the context-sensitive approach to measurement design where predictive accuracy and research feasibility are valued.**V.**

CONCLUSION

This paper has discussed the aspects of selecting the scale used in empirical studies by concentrating on predictive validity implications of single-item and multi-item scales. The comparison reveals that even though multi-item scales have better psychometric strength, single-item measures have an equivalent predictive validity under certain circumstances. The results highlight the relevance of the construct complexity, the aim of the research, and the restraints of practical use in the choice of the measurement instruments.

Practical shortcomings of this research are that it is based on conceptual synthesis and not on primary empirical data and that predictive validity would be different in various disciplines and contexts. Also, the item measures can be affected by the wording of the questions and the interpretation of respondents hence cannot be generalized to the constructs that are complex.

The future course of research must include conducting

massive empirical comparisons in various spheres in an effort to develop more concrete rules on the choice of scale. Future research can examine hybrid methods of measurement, which can integrate the efficiency of single items and the depth of multi-items. The developments in data analytics and machine learning are also open to using the abilities to improve predictive validity testing and scale design. The flexible and evidence-based approach will allow future studies to develop the measurement practices and enhance the overall quality of the empirical research findings.

REFERENCES

[1] Abbott, R. A., Ploubidis, G. B., Huppert, F. A., Kuh, D., Wadsworth, M. E., & Croudace, T. J. (2006). Psychometric evaluation and predictive validity of Ryff's psychological well-being items in a UK birth cohort sample of women. Health and Quality of Life Outcomes, 4(1), 76. https://doi.org/10.1186/1477-7525-4-76

[2] Camisón, C., & Forés, B. (2009). Knowledge absorptive capacity: New insights for its conceptualization and measurement. Journal of Business Research, 63(7), 707–715. https://doi.org/10.1016/j.jbusres.2009.04.022

[3] Castro, M. S., Bahli, B., Ferreira, J. J., & Figueiredo, R. (2023). Comparing Single-Item and Multi-Item Trust Scales: Insights for Assessing trust in Project Leaders. Behavioral Sciences, 13(9), 786. https://doi.org/10.3390/bs13090786

[4] Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. Journal of the Academy of Marketing Science, 40(3), 434–449. https://doi.org/10.1007/s11747-011-0300-3

[5] Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus Reflective Indicators in Organizational Measure Development: A comparison and Empirical illustration. British Journal of Management, 17(4), 263–282. https://doi.org/10.1111/j.1467-8551.2006.00500.x

[6] Graf, L. K. M., Mayer, S., & Landwehr, J. R. (2017). Measuring Processing Fluency: One versus Five Items. Journal of Consumer Psychology, 28(3), 393–411. https://doi.org/10.1002/jcpy.1021

[7] Kamakura, W. A. (2014). Measure twice and cut once: the carpenter's rule still applies. Marketing Letters, 26(3), 237–243. https://doi.org/10.1007/s11002-014-9298-x

[8] Kautish, P., Sharma, R., & Khare, A. (2020). Multi-Item scale development for online consumption emotion construct and psychometric evaluation for relationship

[9] marketing. Journal of Relationship Marketing, 20(2), 91–134. https://doi.org/10.1080/15332667.2020.1717282

[10] Lages, L. F., & Fernandes, J. C. (2004). The SERPVAL scale: a multi-item instrument for measuring service personal values. Journal of Business Research, 58(11), 1562–1572. https://doi.org/10.1016/j.jbusres.2004.10.001

[11] Matthews, R. A., Pineault, L., & Hong, Y. (2022). Normalizing the use of Single-Item measures: Validation of the Single-Item Compendium for Organizational Psychology. Journal of Business and Psychology, 37(4), 639–673. https://doi.org/10.1007/s10869-022-09813-3

[12] McCarthy, J. M., & Goffin, R. D. (2001). Improving the validity of letters of recommendation: An investigation of three standardized reference forms. Military Psychology, 13(4), 199–222. https://doi.org/10.1207/s15327876mp1304_2

[13] Robinson, M. A. (2017). Using multi-item psychometric scales for research and practice in human resource management. Human Resource Management, 57(3), 739–750. https://doi.org/10.1002/hrm.21852

[14] Sarstedt, M., Diamantopoulos, A., Salzberger, T., & Baumgartner, P. (2015). Selecting single items to measure doubly concrete constructs: A cautionary tale. Journal of Business Research, 69(8), 3159–3167. https://doi.org/10.1016/j.jbusres.2015.12.004 Tian, K. T., & McKenzie, K. (2001). The Long-Term predictive validity of the consumers' need for uniqueness scale. Journal of Consumer Psychology, 10(3), 171–193. https://doi.org/10.1207/s15327663jcp1003_5

[15] Varshneya, G., & Das, G. (2016). Experiential value: Multi-item scale development and validation. Journal of Retailing and Consumer Services, 34, 48–57. https://doi.org/10.1016/j.jretconser.2016.09.010a