Current and Future Research In Multi Stage Testing[1]

April L. Zenisky & Michael G. Jodoin

November 30, 1999

[1] Laboratory of Psychometric and Evaluative Research Report No. 370. School of Education, University of Massachusetts, Amherst, MA.

Among the fundamental objectives in test development is the minimization of measurement error. Under a tradition paper and pencil framework, all examinees see the same test form.[2] That is, examinees are presented the same items to measure the construct of interest. Given this constraint, there are three ways to lower measurement error.

First, increasing the length of the test will improve measurement precision. However, this may not be a suitable solution for reasons such as time constraints. Second, the use of more highly discriminating test items is another method to improve measurement precision. Unfortunately, the construction of more highly discriminating items is not likely to be a practical solution since item construction is both a difficult and expensive task. Third, measurement error near a particular ability of interest may be reduced by increasing the number of items with difficulty near the ability of interest. In criterion-referenced testing, where the ability of interest is a constant (e.g., a classification cutscore), this may prove to be an acceptable solution if sufficient items near the ability of interest are available in the item pool. In norm referenced testing, where the ability of interest is a variable (e.g., the ability of each examine), this is not a tenable solution.

Until the development of Item Response Theory (IRT), the aforementioned three methods for improving the measurement precision of tests

[2] Examinees could also see parallel forms of the tests which would require strict content and statistical equivalence.

were justified with the classical test theory (CTT) concept of test variance. The development of IRT not only provided another framework to justify the use of these procedures to reduce measurement error with the concepts of item and test information, but also provided a mechanism to assign comparable scores to examinees from tests consisting of items that were not strictly statistically equivalent. Along with the development of the computer which assisted with the extensive computations required in this framework, this spurred the development of new testing strategies which used items that are *tailored* or *adapted* to examinee ability.

The focus of this paper is a particular form of adaptive testing that appears to have potential for Microsoft's consideration as a test design in their delivery of credentialing examinations, known as multi-stage testing (MST). Before discussing specific examples of MST, we will attempt to define a nomenclature and general model that will place MST in broader context of computer based testing (CBT).

MST in the Context of Computer Based Testing

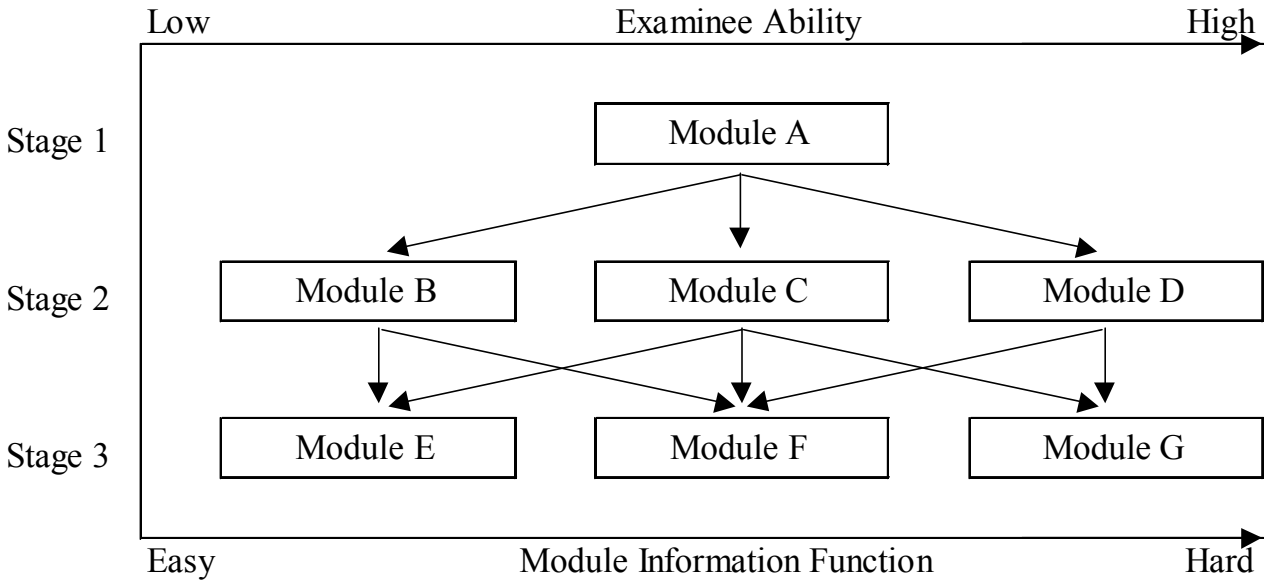
MST is an adaptive strategy with essentially four steps. First, an examinee is presented a set of items called a *module*. A module is a self-contained, carefully constructed, fixed set of items that is exactly the same for every examinee to whom it is administered. Second, the examinee's ability is estimated

based on the items presented. This may be achieved with a variety of methods including maximum likelihood or Bayesian estimation. Third, the examinee progresses to the second *stage* of testing and is assigned a new set of items – a second module - contingent upon the examinees estimated ability level from the previous module(s). Again, the module consists of a predetermined group of items but in the second (and subsequent) stages all examinees will not receive the same module. For example, an examinee with a high ability estimate may be assigned a module consisting of more difficult items, or in IRT terms, a negatively skewed test (module) information function. In contrast, a less capable examinee is presented a module with less difficult items or a positively skewed test information function. In both situations, the module information function should ensure adequate information across the ability levels of examinees that will be assigned to that module.

Finally, steps two and three are repeated for each stage in the MST design. Figure 1 is an example of a three-stage test with three modules at both the second and third stages of testing representing the channeling of low, moderate, and high ability examinees to modules which correspond to their ability level. It is important to emphasize that at the second stage an examinee is administered only one of modules B, C, or D corresponding to low, medium, and high ability examinees, respectively. Similarly, at the third stage examinees are administered only one of modules E, F, or G with allowances made for examinees to be

assigned to more difficult or less difficult modules if ability estimates change after the administration of the second stage module. Finally, the set of modules including the preassigned routing through the stages is referred to as a *panel*. A panel is analogous to a test form in traditional test construction and is useful in test security to avoid excessive numbers of examinees writing identical tests.

Figure 1. A three stage MST.



Linear on the Fly Tests (LOFT) and Computer Adaptive Tests (CAT), two other test designs being considered by Microsoft can be thought of as special cases of MST. Figure 2 demonstrates that a LOFT test may be viewed as a MST

with only one stage but multiple panels. Figure 3 portrays CAT as a *n*-stage MST where each module consists of only one item and *n* corresponds to the number of items need to reach the stopping rule. Although not shown in Figure 3 one could imagine the numerous panels that would accompany a CAT as a result of item exposure controls.

Figure 2. LOFT as a special case of MST.

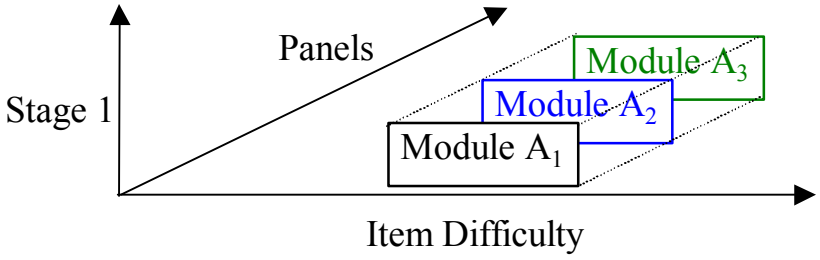
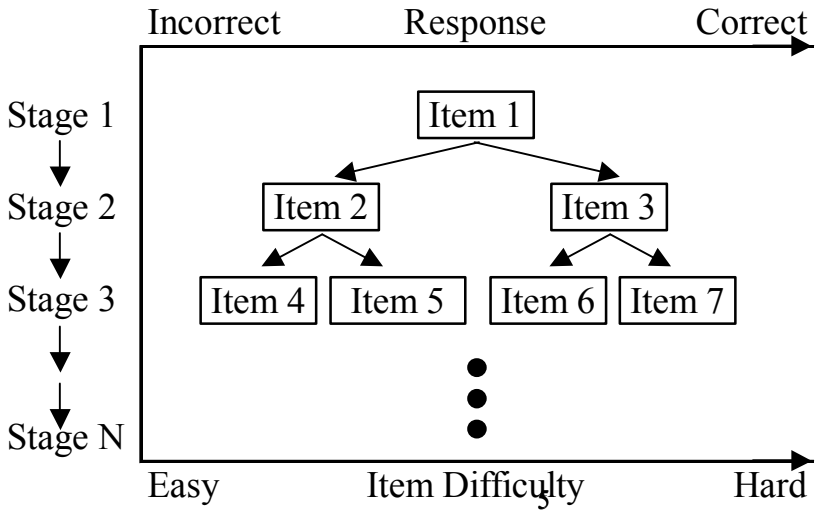


Figure 3. CAT as a special case of MST.



Therefore, one might view MST as a compromise between traditional linear testing methods where all examinees see the same items (or parallel forms for security reasons) and tests that are fully-adaptive at the item level. That is, MST is a partially-adaptive version of traditional linear tests and thereby retains many of the advantages of linear testing while benefiting from the increased measurement precision available with fully-adaptive CAT. Before a discussion of the advantages and disadvantages of MST in comparison to linear and CAT testing, two popular MST designs will be examined.

Two-Stage and Multi-Stage Testing

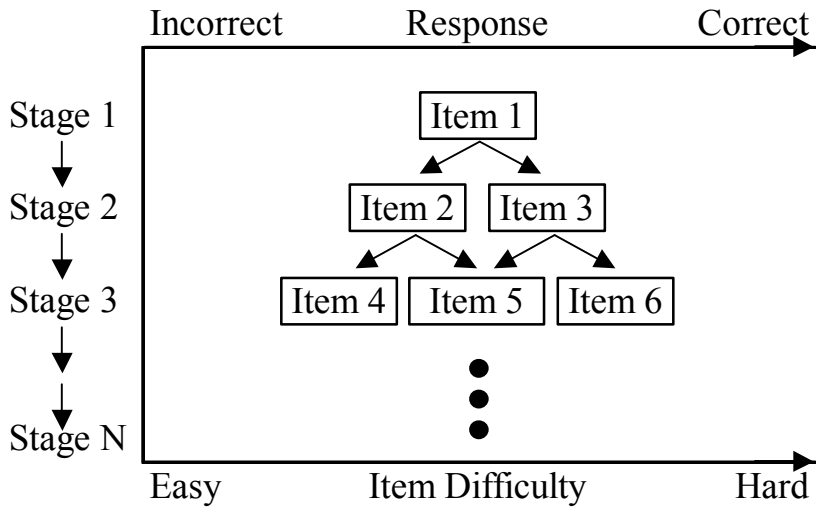
Several test construction strategies incorporating hierarchical structures have been studied and in some cases put into practice, and such strategies may be separated into two-stage and multi-stage approaches (Hambleton, Zaal, & Pieters, 1991). In both the two-stage and multi-stage methods, examinees first take what is generally termed a *routing test* (i.e., the first module), where all examinees respond to some predetermined number of common items. The next set of test items (i.e., the *second stage or module*) an examinee is presented with is determined by the examinee's performance on the routing test. The relative ease of implementing the two-stage test design to groups of examinees or to individuals even without the use of computers has made this format somewhat

more commonly employed than a test structure with multiple stages (Patsula, 1999).

In a MST design incorporating more than two stages, how an examinee performs on the second-stage test then routes the examinee to a third stage module. This process could continue to some programmed level of measurement precision or through a pre-specified number of stages. While some applications of multi-stage testing (MST) used in practice have three or more such stages, most others do employ a two-stage structure (Rock, Pollack, & Quinn, 1995; Luecht, et al., 1996; Rock, 1996). An examinee's ability can then estimated on the basis of the items taken at each of *n* stages.

In both two-stage and multi-stage models, the routing decisions that occur at the end of each stage can be classified as fixed-branching (Hambleton, et al., 1991). Fixed-branching occurs when the structure of items is the same for all examinees, although examinees may move through the structure in a unique way. Figure 4 presents an example of fixed branching where an examinee could end up at item 5 in two ways. Examples of fixed-branching in MST include the flexilevel test suggested by Lord (1971) and Weiss' (1982) stratified-adaptive test. It should be noted that the earliest incarnations of these test structures generally have a single item in each stage, rather than a set of items as is found in designs that are more recent.

Figure 4. A Fixed Branching Example.



The flexilevel test has been described as pyramidal in structure, with a single item at each difficulty level (Hambleton, et al., 1991). An examinee is routed through to different items higher or lower in difficulty according to performance. The items in a stratified-adaptive test are grouped into strata by difficulty, and the examinee is moved through the test by answering an item, being branched into a strata of items based on performance to that previous item, responding to an item in the new strata, and so forth. This process continues up and down across strata until to some level of measurement precision is reached.

It is important to note that multi-stage testing (MST) can be applied in both pencil-and-paper (P&P) and computer-based testing (CBT) formats. Indeed,

Cronbach and Glesser (1965), Lord (1971) and Weiss (1982) introduced P&P methods for adapting tests to examinee ability that can be considered instances of MST. In a P&P administration of a MST, examinees could take the initial routing test, which may be scored by hand using a relatively straightforward scheme such as a number-right scoring, and then be given a second stage on the basis of performance on that routing test. Clearly, this is a less than ideal solution as the scoring involves administrative difficulties in group administrations and fails to take full advantage of the sample independent examinee abilities in IRT. A two-stage or multi-stage test could also be presented to examinees via computer in order to more fully incorporate the advantages associated computer based testing including testing on demand. Given the availability of user-friendly and inexpensive desktop computers, computer based MST has become an economically feasible option. However, practical implementation entails the solution of the unique problem of assembling multiple modules and panels in an efficient manner.

Although the selection of items, from an item pool, to form modules by hand is technically possible even with a large number of content restraints and the fulfillment of other qualitative and quantitative features it is not economically efficient since it requires a lot of time even for a highly trained test developer. This process becomes even more complicated and difficult when several forms or panels are required for computerized administration. Thus, a focus of the MST

literature has been the development of algorithms to assemble tests (Armstrong, Jones, Li, & Wu, 1996; Luecht, 1998; Luecht & Hirsch, 1992; Luecht & Nungester, 1998; Stocking, Swanson, & Pearlman, 1993; Swanson & Stocking, 1993; van der Linden, 1994; van der Linden & Adema, 1998; van der Linden & Boekkooi-Timminga, 1989). Ideally, these automated test assembly algorithms not only need to be flexible enough to develop modules for various MST designs but also should be capable of creating multiple panels that control the overlap of items or modules between panels. This should enable more efficient test development because it would enable each module to be carefully reviewed by test developer for the qualitative features that are not easily achieved by computer algorithms such as gender/racial sensitivity reviews.

A prominent example of an automated test assembly algorithm is Computerized-Adaptive Sequential Testing (CAST; Luecht and Nungester, 1998). CAST automatically assembles modules based on prespecified test (module) information functions. In addition, it allows for a variety of MST designs or panel structures including numerous content constraints. Furthermore, it provides an ability to control the item overlap between panels. Luecht and Nungester (1998) provided an example where a three stage MST panel with 426 content restraints was assembled in less than 20 seconds with their CAST software. In a related example, 99% of the 588 content restraints were satisfied for a comparable three stage MST design using the CAST software with the limitations being set by the

quality of the item pool. Clearly, such automated test assembly programs will be necessary tools in the efficient development of a computerized multi-stage based testing program.

Advantages and Disadvantages of Computerized MST

In large part, both the advantages and disadvantages of MST are derived from its intermediary position between LOFT and CAT. That is, MST acquires some of its advantages from the desirable properties it has in common with CAT that, in turn, are the less desirable properties of LOFT. Similarly, MST obtains some advantages from the desirable properties if has in common with LOFT which are considered disadvantages of CAT. These include measurement precision, parsimonious design, test form quality control, better use of the item pool, and the opportunity for item review by examinees.

A frequently noted benefit of a CAT is enhancement in measurement precision for a fixed number of test items. This creates the opportunity for potentially shorter and more efficient tests. This is also the critical disadvantage to linear testing. Since MST may be considered partially-adaptive, it shares this key feature with CAT and ameliorates this weakness in LOFT. Within a panel structure, the selection of a module to be administered at a particular stage is dependent on the performance of examinees relative to previous stages. That is, the between-stage adaptation to match examinee ability to modules consisting of

items with a corresponding narrow range of item difficulties results in more precise ability estimates. Subsequently, this may be used to increase testing efficiency or shorten tests as compared to LOFT tests with comparable item pool quality. Conversely, MST results in less ability estimate precision in comparison to CAT since there are fewer opportunities for adapting to examinee ability estimates (Patsula, 1999). However, due to item exposure controls to increase test security and imperfect ability estimates particularly early in CAT, the adaptation in CAT is less than optimal and may reduce the differences in efficiency between MST and CAT. In fact, Luecht, Nungester, and Hadadi (1996) reported comparable MST and CAT ability estimate precision. In high-stakes credentialing and licensure examinations such as at Microsoft, however, the precision of the ability estimates is of critical importance, and so this ability estimation aspect of MST remains an area deserving of further research with a variety of MST designs.

A second benefit of MST is a simplification in test structure (Schnipke & Reese, 1999). Aside from the obvious concerns for implementation, a critical issue for test developers considering a CAT design involves the difficulty in explaining item selection and scoring methods. Although typically based on the similarly complex algorithms, the concept of stages and routing based on performance for a group of items or module may seem more straightforward and intuitive to examinees and other stakeholders.

An additional benefit of MST is the capacity to carefully craft and control the quality of all possible test forms (panels). Historically, test construction has been a subtle and nuance laden task which might be aptly described as an art as much as a science. In contrast, a CAT is actively assembled item by item via an item selection algorithm as the examinee responds to each item. Although programming a CAT to meet numerous test specifications is not an impossible task, incorporating more qualitative constraints such as gender balancing and item interdependence are significant concerns for test developers. The structure of MST allows for the construction of numerous modules that could be carefully reviewed in advance of the test administration. That is, as sets of items for each stage are constructed beforehand allowing for a level of quality assurance that is not possible in CAT to be introduced to the test construction process (Patsula, 1999).

This is not to suggest that in CAT items in the item pool are arbitrarily selected and retrieved for use but rather to emphasize that MST provides a reasonable opportunity for careful consideration of all item sets that an examinee may be administered. In addition, the partially-adaptive nature of MST affords the test developer the opportunity to retain the increased accuracy in measurement and at the same time better meet content and cognitive skills specifications of a test (Reese, Schnipke, & Luebke, 1999). As Reese, et al. (1999) pointed out, MST could prove especially valuable in addressing this concern, as the pre-

assembly of the modules may facilitate the meeting of such requirements and thus lead to more valid tests.

Finally, pre-assembly of modules in MST may permit the opportunity for test developers to take greater advantage of the item pool, a very real concern in many CAT applications. In pulling together item sets for presentation in various stages, a multitude of combinations of discrete items could be created, allowing for the conscious attempt to administer a wider sampling of items into modules than sometimes occurs in the selection of discrete items in CAT.

Another key advantage of MST is the ability to allow item review. A traditional feature of test taking that examinees are accustomed to is the ability to go back and review their responses. Since ability estimates are made after each item is administered and subsequent items depend on this estimation, this feature is not available in CAT. For examinees, this is a primary source of anxiety with the format and perhaps its greatest challenge for more complete public acceptance (see, for example, criticisms of the GRE). In MST, the opportunity to review the items within a module would not have an impact on the psychometric properties of the test and therefore could be allowed.

Finally, a cited potential benefit of MST is the minimization of item exposure (Luecht, et al., 1996; Patsula, 1999). These authors argue that by controlling the amount of overlap between panels and the number of panels in use that security risks analyzed and controlled. However, there has been little

discussion of remedies when items are overexposed, memorized, or in some other way inadvertently made public. In both linear and CAT, the solution may be as relatively simple as removing a test form or the exposed items from the available item pool. In contrast, the problem may be more severe in MST as an item could be integrated in several panels as a part of module at various stages. Simply removing the panels with exposed items could result in a large reduction in the available panels and present serious security concerns. This would be especially troubling since each panel, presumably, would have been carefully reviewed as this is a fundamental advantage of MST over CAT. It may be MST contains the security disadvantages of both linear and CAT.

Areas for Future Investigation

More research needs to be done in the area of applying MST to credentialing examinations, as this is the primary focus for Microsoft. Much of the work that has been completed has focused on the precision of ability estimates across the ability continuum, rather than around a specified cutoff score where a credentialing decision will be made. To this end, issues that remain to be studied include concerns in the area of test design and test security.

First, the number of stages that should be included in a MST to balance measurement precision and design parsimony should be addressed. Although research has shown MST to be more efficient than linear and less efficient than

CAT (Luecht, Nungester, & Hadadi, 1996), the nature of this compromise should be carefully examined to guide decisions on the appropriate numbers of stages for various applications. In particular, decision consistency rather than standard errors of measurement need to be considered (i.e., criterion-referenced rather than norm-referenced applications).

Second, the design of each stage should be considered. Lord (1980) identified several points including the number of items, and the overlap and specification of target module information functions to be considered in the process of developing a two-stage test. Patsula (1999) extended this research to three-stage tests. Continued research in this area is warranted for larger MST and mixed MST/ CAT designs. Of importance is the number and nature of items included in the routing test for the best initial ability estimates (Patsula, 1999) and the implications of this for decision consistency.

Third, security concerns deserve careful thought and research. The effect of publicly exposed items on ability estimates and defensible mechanisms for dealing with such events need to be developed. This should include additional work on mechanisms to insert and remove items into modules and panels both as a matter of regular item pool maintenance and as a result of unexpected events. Specifically, mechanisms are necessary to enable a reduction in the exposure of items near the cut point, as these items are likely to be frequently utilized in credentialing applications.

References

Armstrong, R. D., Jones, D. H., Li, X., & Wu, I-L. (1996). A study of a network-flow algorithm and a non-correcting algorithm for test assembly. *Applied Psychological Measurement*, *20*, 89-98.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions, 2nd Ed*. Urbana: University of Illinois Press.

Hambleton, R. K., Zaal, J. N., & Pieters, J. P. M. (1991). Computerized adaptive testing: Theory, applications, and standards. *Advances in educational and psychological testing*. Boston, MA: Kluwer Academic Publishers.

Lord, F. M. (1971). The self-scoring flexi-level test. *Journal of Educational Measurement*, *8*, 147-151.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Luecht, R. L. (1998). Computer-assisted assembly using optimization heuristics. *Applied Psychological Measurement*, *22*, 224-236.

Luecht, R. L., & Hirsch, T. M. (1992). Item selection using an average growth approximation of target test information function. *Applied Psychological Measurement*, *16*, 41-51.

Luecht, R. L., & Nungester, R. J. (1998). Some practical examples of computerized-adaptive sequential testing. *Journal of Educational Measurement*, *35* (3), 229-249.

Luecht, R. L., & Nungester, R. J. (1996, April). *Heuristic-based CAT: balancing item information, content, and exposure*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Patsula, L. N. (1999). A comparison of computerized-adaptive testing and multi-stage testing. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Reese, L. M., Schnipke, D. L., & Luebke, S. W. (1999). *Incorporating content constraints into a multi-stage adaptive testlet design*. (Law School

Admissions Council Computerized Testing Report 97-02). Newtown: PA: Law School Admissions Council.

Schnipke, D. L., & Reese, L. M. (1999*). A comparison of testlet-based test designs for computerized adaptive testing*. (Law School Admissions Council Computerized Testing Report 97-01). Newtown: PA: Law School Admissions Council.

Stocking, M. L., Swanson, L., & Pearlman, M. (1993). Application of an automated item selection method to real data. *Applied Psychological Measurement*, *17*, 167-176.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*, 151-166.

van der Linden, W. J. (1994). Computerized educational measurement. In T. Husen and T. N. Postlethwaite (Eds.) *International Encyclopedia of Education (2nd Ed.)* (pp. 992-998). Oxford: Elsevier.

van der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, *35* (3), 185-198.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A minimax model for test design with practical constraints. *Psychometrika*, *54*, 237-247.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24* (3), 185-201.